Universidad del Norte
Department of Electrical and Electronics Engineering

Intelligent system for non-technical losses management in electricity users

By
Rubén Darío González Rodríguez

MASTER THESIS

Advisor
Dr. Christian G. Quintero M.

Barranquilla, Atlántico, Colombia
June 2017

Universidad del Norte
Department of Electrical and Electronics Engineering

Intelligent system for non-technical losses management in electricity users

A dissertation presented to the Universidad del Norte in partial fulfillment of the requirements for the degree of MASTER OF SCIENCE

By

Rubén Darío González Rodríguez

Advisor:

Dr. Christian G. Quintero M.

Barranquilla, Atlántico, Colombia
June 2017

Universidad del Norte
Department of Electrical and Electronic Engineering

ABSTRACT

Intelligent system for non-technical losses management in electricity users

By Rubén Darío González Rodríguez

Advisor: Dr. Christian G. Quintero M.

The central axis of this research is the problem associated to non-technical losses of electrical energy that are the product of the fraudulent behavior of the electrical system users. In the particular case of the Colombian Caribbean coast, the region's retailer reports that only in 2015 the losses due to theft of energy were 17.23% of the total energy purchased by the company. In addition, the company reports that for the same year the effectiveness rate of the energy recovery plans implemented was 68%. The main effort of this thesis is to propose a methodology that allows to increase the effectiveness rate of the energy recovery process. The way to achieve this objective is by characterizing the variables that allow to model the behavior of a user of the electricity sector, so that according to the model it is possible to classify fraudulent and non-fraudulent users.

An intelligent system for the management of non-technical losses associated to the fraudulent behavior of users is proposed. The above translates into a set of algorithms that receive as input the variables characterized and based on computer intelligence techniques (machine learning), allow to cluster users through their different consumption profiles, modeling such profiles to predict

their future behaviors and classifying fraudulent and non-fraudulent users. The implemented system consists of three (3) stages, a non-supervised clustering of consumption profiles based on a hybrid algorithm between self-organization maps and genetic algorithms, a future consumption prediction based on ARIMA and intelligently corrected through a neural network, the final stage is a binary classifier based on random forests.

The proposed intelligent approach was trained and tested in a real case of study constituted by the database of clients belonging to the cities of Barranquilla, Puerto Colombia and Malambo. The results allow to demonstrate that an improvement in the process of detection of fraudulent users was obtained, significantly increasing the index of effectiveness compared to non-intelligent approaches and intelligent approaches, both previously applied by the region's retailer.

Acknowledgments

First, I want to thank God for allowing me to fulfill this dream, to give me health and fill me with intelligence, wisdom, patience, constancy and all those gifts that helped me to develop this work satisfactorily. I know that without him in my life none of this would have been possible and I hope he allows me to live to fulfill all the desires of my heart.

To my father Carmelo González and my mother Angela Rodríguez, I want to thank you for always believing in me, for accompanying me during the whole course of my master's studies and supporting me in all the aspects that were necessary. Thank you for teaching me that with effort and dedication it is possible to achieve each of the goals that we draw no matter how difficult they seem.

To my girlfriend Tatiana Rodríguez for traveling with me all this way, for accompany me in each of the days, for filling my life with her love and always give me her unconditional support. Thank you for helping me in every situation and being involved in every decision, for always wanting the best for me and reaching with me every academic achievement in my life.

I want to especially thank my advisor Christian G. Quintero M. for giving me the opportunity to work with him, for believing in my abilities and helping me to get this project forward. I am immensely grateful for the time he has invested in me and the help I have received from him as an advisor, teacher and friend. Thank you for helping me grow as a professional and researcher and for teaching me that the important thing is not the problems but the solutions.

To my co-advisor Jose Oñate and Jamer Jimenez for always being willing to collaborate, for allowing me to learn from and with them, to help me in the development of each of the stages of this project and to serve as guides at the

academic and personal level. I am sure the road would have been more difficult without the presence of them during this time.

Thanks a lot to Katheryn Donado, Daniel Turizo and José Cayón, people that I had the pleasure of knowing during the masters course and with whom I shared academic and personal experiences that made us very good friends. Best wishes to them in their personal projects.

Likewise, I would like to thank those members of the Department of Electrical and Electronic Engineering who provided me with academic, professional and personal support during the Master's Degree.

Finally, I would like to thank all those people who were not named but somehow contributed to my development and growth during this time. Without their friendship, support and help I would not have been able to achieve this goal.

Dedicated to my God, my parents, my love and my family

General Contents

Detailed Contents

List of Figures

List of Tables

PART I

INTRODUCTION AND RELATED WORK

Chapter 1

Introduction

This chapter provides an introduction to the work presented in this thesis. Specifically, the motivation in the research area, the pursued aims and the main contributions are briefly described. Finally, the chapter concludes with an overview of the structure and contents of the thesis.

1.1. Motivation

As part of the real world processes, power systems are not capable of delivering to users all of the electric power produced at power plants, so inherently to their operation there will always be a difference between the energy generated and delivered to end users. In a technical way, this difference is defined as a loss of energy (Navani, Sharma & Sapra, 2012) because it is not remunerated in any way, and in turn has an environmental and economic impact on the operation of electric systems. Since this energy must be generated and therefore will require a greater use of primary sources and fossil fuels in generation plants, causing overcosts in the operation of the electricity network.

Losses of electrical energy can be classified into two large groups, technical losses and non-technical losses. The former correspond to the percentage of electrical energy that is transformed into other types of energy during the transmission process, which is a product of the characteristics of the materials with which the different equipment that compose the system are built. The second group refers to the amount of electrical energy that is delivered to end users but is not billed properly and therefore is not economically represented. Within this group of non-technical losses are those products of errors and breakdowns in the measurement equipment, energy not invoiced due to errors in meter reading or billing processes of trading companies and finally those associated to fraudulent behavior of

customers, who in one way or another evade the payment corresponding to the consumption of electric energy (Suriyamongkol, 2002).

The development of this master thesis will focus on the non-technical losses of electrical energy that are the product of fraudulent behaviors of the clients. According to the Unidad de Planeación Minero Energética (UPME), this type of losses can be assumed on average as 15% of the total energy purchased by the country's electricity commercialization companies (Unidad de Planeación Minero Energética, 2011) (see figure 1.1-1). This represents a warning value due to the economic and environmental impacts that these losses bring with it and therefore, it is important to propose solutions that are aimed at reducing these high rates of electric energy losses.

Fig 1.1-1. Energy losses in distribution systems. [3]

Currently, electrical energy commercialization companies are implementing improvement and management plans to reduce the percentage of energy lost. This has been reflected in the statistics that describe a decrease in the total number of them since 1998, where the percentage of non-technical losses was 27% (Unidad de Planeación Minero Energética, 2011). Among the solutions currently implemented for the detection of fraud, three types stand out: consumption deviation, commercial cycle evaluation and macromeasuring. The consumption

deviation consists in quantifying the total energy losses in a specific zone and dividing that value equally among the number of users of the same. Then it is analyzed whether the amount of energy that could be recovered by inspecting each user of the area represents a profit for the company, taking into account the expenses associated with sending crews to check each house. The commercial cycle is a follow-up that is done to customers who have presented drastic changes in their consumption, turning them into potential cases of irregularities. For these customers are evaluated the behavior of consumption in previous months and the same month where the change occurred but in previous years, with the purpose of obtaining information that allows to conclude in a safe case of fraud. Finally, and the most used at the moment, is the macromeasuring, technique that consists in the location of meters in the output of the transformers that feed the groups of users, which in case of no fraud exists the measurement delivered by these meters must coincide with the sum of the measurements of the clients associated with that transformer, otherwise it is verified that there are losses of energy in that group of clients.

In spite of the above, the amount of lost money that the commercialization companies report related to the fraudulent connections continues to be alarming. This due to the increasing development of new ways ranging from craft methods to application of technical knowledge for avoidance of payment for the consumption of electrical energy. So much has been the impact of this problem that is currently being the focus of multiple worldwide research that converge towards the use of algorithms and computational intelligence techniques. Such techniques allow the exploration, acquisition, processing and analysis of large amounts of data that any human being operator would be incapable, as well as a high degree of precision in the detection of frauds which would allow recovering a large amount of the money associated with this type of losses. In addition, it is looked for the decrease of the money that is invested in the sending of visiting crews that result in failed inspections.

Some authors who have tackled this problem from the perspective of intelligent algorithms and machine learning have been able to make significant contributions.

Having obtained satisfactory results that confirm that this type of techniques become powerful tools which provide better solutions than those currently used. In turn, the success of these works encourages that the search for better and better solutions to this problem continue to be from the approach of computational intelligence techniques, as is the case of this research.

Several computational intelligence techniques have been used to propose solutions to the problem of fraudulent connections: neural networks (Czernichow, Muñoz & Sanz-Bobi, 1998; Cabral, Martins, Pinto. A, & Pinto. J, 2008; Biscarri. F, Biscarri. J, Guerrero, León, Millán & Monedero, 2010), support vector machines (Hashim, Hussien, Mohamad, Pok & Yak, 2007; Nagi, Yap, Tiong, Ahmed & Mohammad, 2008; Nizar & Dong, 2009) and fuzzy logic (Figueiredo, Muniz, Tanscheit & Vellasco, 2009; Carmona, Nunes, Saavedra & Silva, 2011). This allows to demonstrate the interest that exists in the use of these tools to obtain more rigorous and effective solutions to the problem of non-technical losses due to fraudulent connections.

In this research, a solution to this problem is proposed by implementing an intelligent system to detect users with fraudulent connections divided into three stages. The first is a clustering of consumption profiles based on a hybrid algorithm between neural networks and genetic algorithms. This clustering stage was developed taking into account the consumption curve of each user. The second stage is a predictor of future consumption based on ARIMA models and corrected by a neural network. This prediction stage was developed taking into account the customers consumption curves and exogenous variables such as temperature, billed days, socio-economic stratum, among others. Finally, the final stage is a binary classifier of fraudulent and non-fraudulent users based on random forest, which receives as inputs the variables that characterize the consumption behavior of each user, as well as the outputs of the two stages mentioned above.

1.2. Objectives

The research conducted in this dissertation is oriented to the development of intelligent clustering, prediction and classification algorithms, which, knowing the

variables that characterize the consumers of electric energy, together with those exogenous variables that affect consumer behavior are used to improve the process of detecting users with fraudulent connections.

- **Problem**: Detection of users fraudulently connected to the distribution network to improve the results of the energy recovery campaigns and therefore, reduce the rates of non-technical losses of electricity.
- **General Objective**: Develop and implement an intelligent system for power consumption behavior analysis and detection of non-technical energy losses caused by fraud in the electrical system users.
- **Goals:**
  - ✓ Characterize the most relevant variables for energy consumption behavior analysis. These should allow clustering, prediction and classification of consumption in electrical energy users, so that may be possible to detect fraudulent behavior in these users.
  - ✓ Design and implement the intelligent system for consumption analysis and detection of non-technical losses taking into account the variables above characterized.
  - ✓ Evaluate the performance of the intelligent system as a tool for detecting electrical frauds.

1.2.1. Thesis Question

The principal question addressed in this dissertation is:

¿Could a computational intelligence system be a tool to improve the detection process of users with fraudulent connections?

1.2.2. Approach

In this research is proposed the design of a methodology that allows the detection of fraudulently connected users to the distribution network. The above through the integration of algorithms of unsupervised clustering of consumption profiles, prediction of future consumption and classification of users in fraudulent and not fraudulent. This proposal seeks that the detection be carried out by an intelligent system, that is to say, making use of algorithms that are characterized by learning

through experience, which have demonstrated that are able to perform this type of tasks with performances much superior to the obtained with human operators. This approach differs from those presented in other works because it seeks to reinforce detection by including additional variables that are the results of other intelligent algorithms.

Computational intelligence techniques belong to the branch of computer science that is known as machine learning, a field in which the raw material around which everything turns is information. Machine learning algorithms are able to move volumes of information represented in millions of data, an amount that far exceeds the processing capacity of any human being. From the above, it is possible to infer that to be able to apply this type of techniques in the solution of a problem it is necessary to guarantee the quality of the information on which will be worked. Therefore, in order to fully comply with the above, the proposed system has an initial stage of conditioning in which filtering processes are carried out, incomplete and erroneous data are eliminated, atypical data are adjusted and it is verified that for each variable exists the total of the information.

Referring specifically to the methodological scheme proposed in this research, the starting point are the variables that allow to characterize the electric power consumption behavior of each user. This point is common for all approaches proposed in other works since the data corresponding to the values of these variables represent in all cases the inputs to the system. The difference of the proposed approach is that in addition to the aforementioned inputs, some of them are used for the execution of two stages whose outputs accompany the set of initial inputs of the system, thus allowing to reinforce the detection of fraudulent users.

The first of the two stages described above corresponds to an unsupervised clustering of electrical energy consumption profiles. This algorithm is responsible for dividing the universe of users that will handle the system in different groups, taking into account that those users belonging to the same group must present common consumption behaviors. The reason for the implementation of this stage is that within each group will be captured all types of features that are shared by

users within it, this would include among many other patterns of fraudulent behavior in such users.

The second stage is a prediction of future consumption of each user. This algorithm is responsible for modeling the consumption curve of each user, so that it is possible to make predictions of its future values. The reason behind the implementation of this algorithm is due to being possible to completely model the behavior of a user, it is also possible to make estimates of future behavior. If there are significant differences between the predicted values and those actually occurring in those periods, there would be reasons to suspect about possible abnormal behaviors in those users.

Finally, the fraudulent users detection stage is a supervised classification algorithm, which receives as input all system variables together with the two additional inputs that are the product of the intelligent steps described above. This algorithm is trained to be able to distinguish between users with fraudulent behaviors and those with non-fraudulent behaviors. The inclusion of the two additional variables is the key difference between this proposal and others. Such variables provide information of a considerable weight that allows reinforcing the learning, making possible that the algorithm makes a better characterization of the fraudulent and non-fraudulent behaviors.

1.3. Contributions

This thesis makes the following contributions in problems about management of non-technical energy losses related to fraudulent behaviors:

- *An intelligent approach for detecting users with fraudulent connections. Starting with the selection of the variables that characterize the behavior of electrical energy consumption, then the stages of clustering of consumption profiles and prediction of future consumption are established. The above in order to obtain an additional set of inputs that accompany the initial variables in the final stage which is the detection. Those two stages will allow to improve the results of the process by means of a reinforcement in the training of the classifier (detector). The clustering of profiles is based on a hybrid algorithm*

between genetic algorithms and self-organized maps, the prediction consists of a statistical method that adjusts the best ARIMA model of each user to later correct it through a neural network, the final stage is a binary classifier based on random forests.

- *To improve the process of detection of fraudulent users, the algorithms that compose the stages of the computational system were developed. These correspond to the clustering of profiles, prediction of consumption and classification of fraudulent and non-fraudulent users. In order to achieve the correct functioning of the system, data loading and data conditioning algorithms were implemented.*

1.4. Reader's Guide to the Thesis

Following is a general description of the contents of this dissertation. This master thesis is organized in three main parts distributed by chapters.

Part I: Introduction and Related Works

Chapter 1 presents a motivational introduction on the main topics, objectives and contributions regarding this dissertation.

Chapter 2 gives a general overview of background information regarding Non-technical losses of electrical energy, fraudulent connections in the distribution network and all those theoretical concepts associated to the computational intelligence techniques used, as well as all the mathematical and statistical basis that supports this research, which are required to develop the propose approach described in chapter 4 and 5.

Chapter 3 provides a general survey of the most relevant work related to the research addressed in this thesis.

Part II: Proposed Approach

Chapter 4 describes the formal aspects of the intelligent system for non-technical losses management model presented in this thesis.

Chapter 5 presents the implementation of the approach proposed in chapter 4. The chapter also contributes to complete the description of such proposal.

Part III: Results and Conclusions

Chapter 6 provides experimental results of the implemented approach. The results of the executions are shown separately for each stage of the system, as well as the execution and testing of the system as a whole for the case of study.

Chapter 7 discusses and analyzes the results, summarizes the conclusions and contributions of the thesis and outlines the most promising directions for future work.

Chapter 2

Background Information

General concepts of the production and consumption of electrical energy, technical and non-technical losses of energy, emphasizing the latter because it is the problem to be solved through the development of this research. In addition, the computational intelligence techniques related to the development of the proposed approach will be described.

This chapter is divided into three sections in order to answer three major questions. The first, to which area of study of Electrical Engineering belongs the research developed?, which is proposed with the aim of providing the reader with the bases to know ***where*** the development has been focused. The second, what is the problem on which this thesis is centered?, with which it is tried to define in ***what*** was worked. The last question is how to solve this problem?, which introduces the concepts that allow the reader to understand ***how*** the problem can be solved from the addressed approach.

Concepts associated to the area of study of the Electrical Engineering to which the research belongs

2.1. Production and consumption of electrical energy

2.1.1. Power system

Power system is the grid that results from the interconnection of electrical equipment whose purpose is to generate, transport and consume electrical power. Particular emphasis is placed on the word power because these types of systems are used to provide electricity to cities, regions and even countries, which are loads that demand considerable amounts of electrical energy.

These types of systems may have different topologies, be designed to handle different values of electrical power and even cover different distances. However, generally have a similar structure that allows dividing the cycle of electrical energy into four fundamental stages: generation, transmission, distribution and consumption (Kundur, 1994). This process is illustrated in figure 2.1-1.

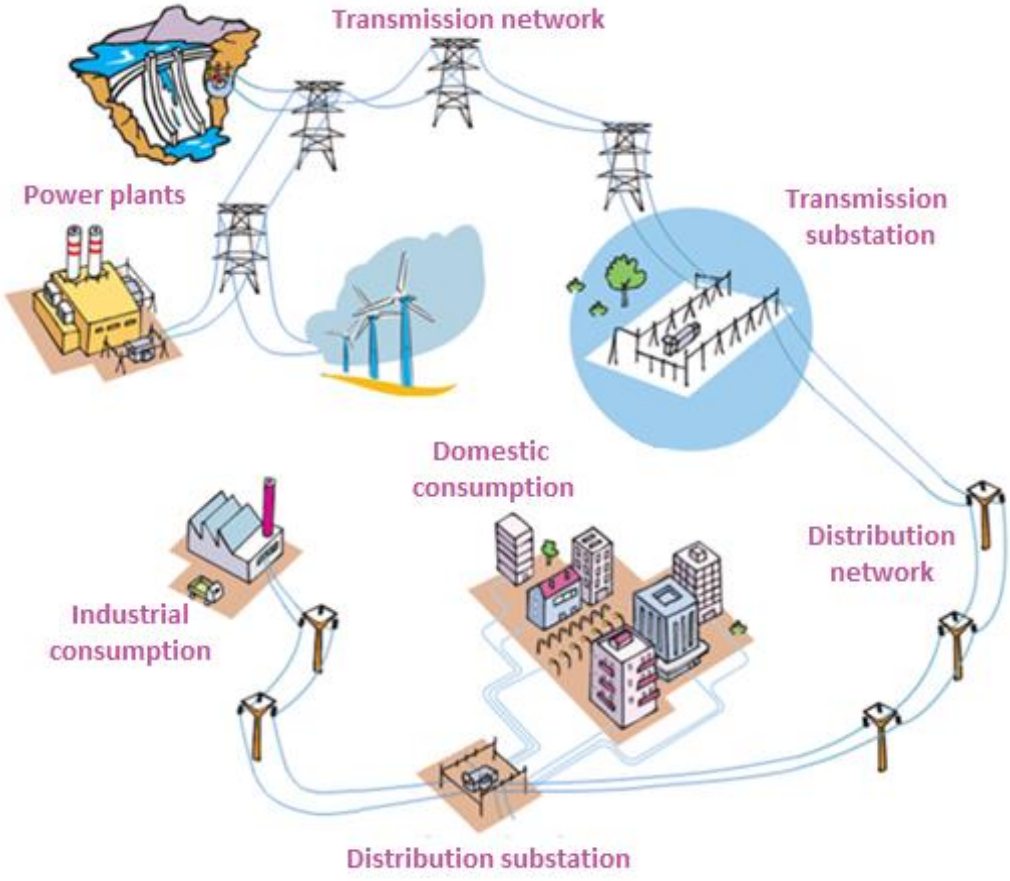


Fig 2.1-1. General structure of power systems. (www.intercolombia.com)

1. Generation is the initial stage of the process and takes place in the generation plants (Fig 2.1-1). In this part, the energy of the primary sources is converted into electrical energy. However, in most cases primary sources are located in remote locations and the electrical energy produced can not be delivered immediately to consumers. This case fits perfectly with the Colombian scenario, where most of the energy generated is obtained from hydroelectric plants located in river reservoirs, and the other part is obtained from thermoelectric plants that require large bodies of water for their thermodynamic processes.

Therefore, the energy obtained must be taken to the urban centers where it will be consumed, giving way to the next stage of the process.

2. The second stage of the process is known as transmission and its purpose is to transport the electrical energy produced in the generation plants to the consumption centers where it will be used by the end users. This task is carried out using a transmission network (Fig 2.1-1). Such network is made up of special equipment that allows the transport of energy with voltage values close to hundreds of thousands of volts, this for the purpose of making this stage as efficient as possible and that the largest possible amount of energy can be brought to the centers of consumption.
3. When the energy is already in the urban centers, it is not possible to consume it immediately, the above because the energy is delivered at points of interconnection of the transmission network known as electrical substations. In these places transformations that allow to reduce the levels of voltage to values suitable for the use of the equipment of the end users are made. In addition, it is necessary to have an infrastructure similar to the transmission network but whose purpose is to deliver the energy to each user of the system, this set of equipment is known as the distribution network (Fig 2.1-1) and corresponds to the third stage of the process.
4. The final stage of the process carried out in a power system is the consumption of electrical energy by each of the end users. From figure 2.1-1 it is possible to observe that the energy consumption is divided according to the type of activities carried out by users, with users mainly types are residential, commercial and industrial in the Colombian case. Usually refers to this stage as commercialization because users regardless of their type of consumption must pay a price for the electric power service.

2.1.2. Types of users of the electrical system

Users in the electricity sector are initially classified as regulated and non-regulated. The regulated ones correspond to those whose rates for the payment of electric energy are established by the state. The non-regulated are users whose

consumption exceeds a quota imposed by the government and, therefore, they are allowed to negotiate with the companies of the electric sector their own tariffs.

The development of this research is located within the group of users belonging to the group of regulated, so that the types of users that make up this group will be defined. Figure 2.1-2 shows an illustrative scheme proposed for this purpose.

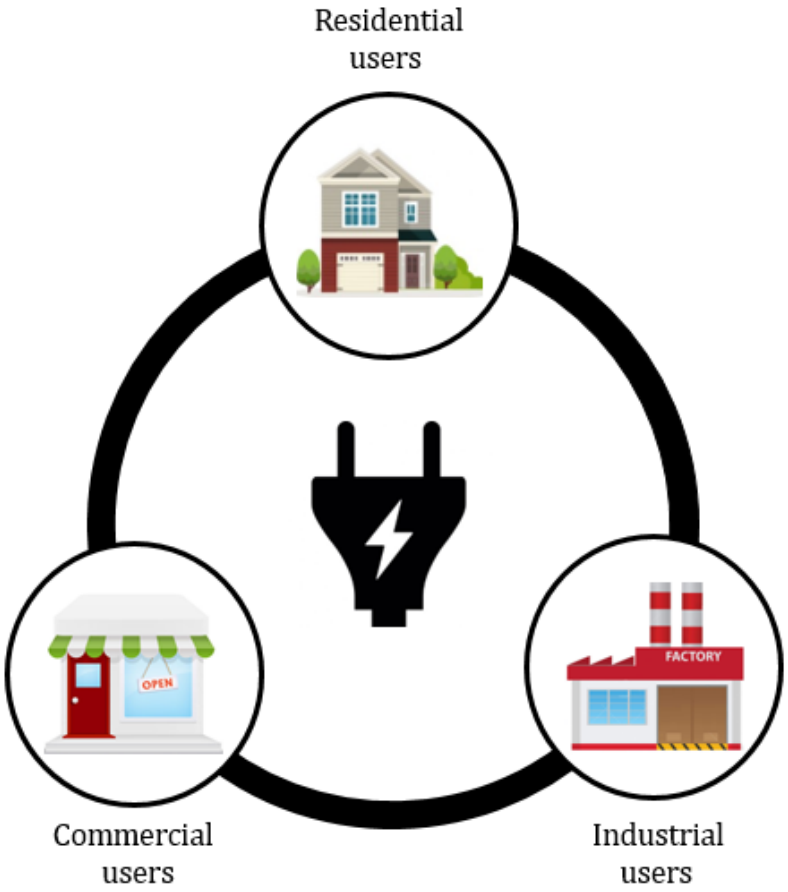


Fig 2.1-2. Scheme of the types of regulated users according to the use of the energy service.

From Figure 2.1-2 it is possible to observe that there are three large groups of regulated users, which are differentiated by the type of activities they perform daily, these activities define their habits of electrical energy consumption.

- **Residential users:** They are divided by socio-economic stratum from 1 to 6. This type of users represent 90% of the energy consumers of the Colombian electrical system. They are characterized because their energy consumption is derived from the domestic activities that are carried out at the family level within the home, which are directed toward the comfort

and quality of life of those who integrate the dwelling (Ley N° 142, 1994; Ley N° 286, 1996).

- **Commercial users:** They correspond to consumers whose use of electric energy is directly related to the performance of some economic activity, in which the customer makes use of the service to operate equipment that allows him to develop a process from which he expects to obtain economic benefits (Ley N° 142, 1994; Ley N° 286, 1996).
- **Industrial users:** the definition of industrial customers corresponds exactly to the definition of commercial customers, in which consumption is related to the use of equipment to develop activities that allow economic benefits to be obtained. The difference between a commercial customer and an industrial customer is based on the amount of energy each consumes for the development of its economic activity, being the consumption of industrial customers considerably larger than commercial customers. Generally, industrial users are associated with mass production processes, in which high electrical consumption machinery is used. Industrial consumers that belong to the group of regulated users are known as small industrialists, this because their consumption, although it is significantly higher than residential and commercial users, does not reach to exceed the quota to be considered non-regulated users (Ley N° 142, 1994; Ley N° 286, 1996).

2.1.3. Electrical energy consumption profile

The electric power consumption profile, also known as the load profile, represents the electricity consumption behavior of a user measured in a set period (Elexon, 2013). Although the consumption profile is a time series, it is usually represented graphically to facilitate its analysis. The most commonly used consumption profiles are constructed using periods of days, months and years. As an example, the daily typical consumption profiles of the Sistema Interconectado Nacional Colombiano (SIN) are shown in Figure 2.1-3.

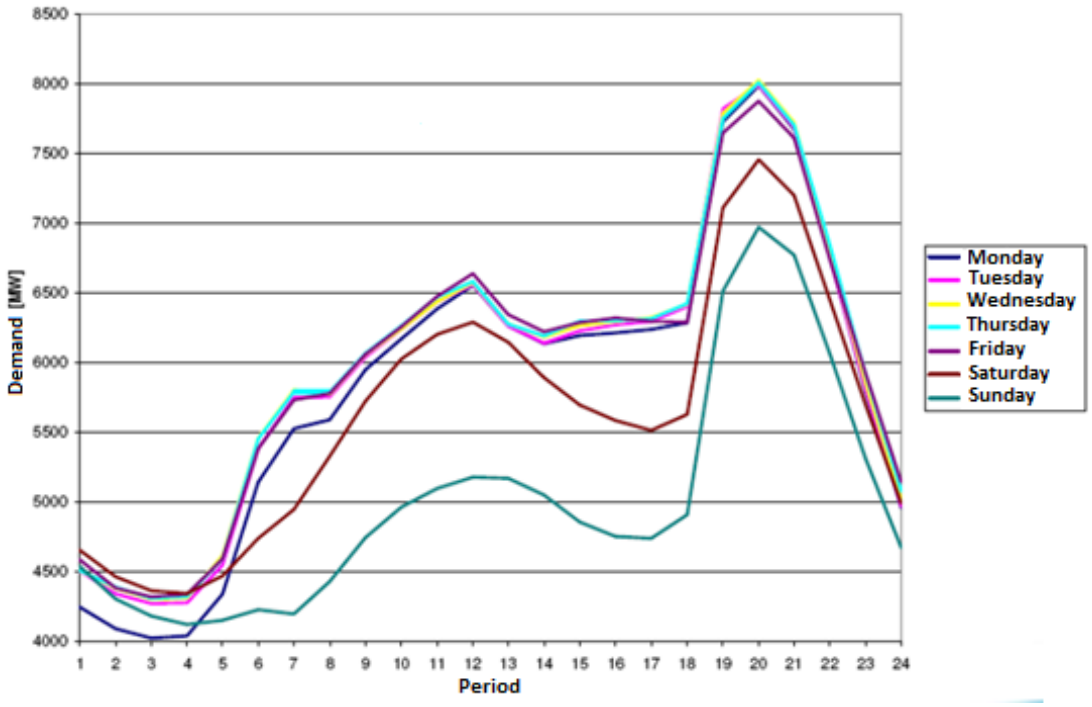


Fig 2.1-3. Daily typical consumption profiles of the SIN. (Derivex, 2010)

Concepts associated with the problem on which this research is focused

2.2. Electrical energy losses

2.2.1. Technical losses of electrical energy

A technical loss is defined as the amount of electric energy transformed into other types of energy during the process of transporting it from the generation plants to the centers of consumption. These losses occur due to the physical properties of the materials and are inherent to the process of electric charges circulation in conductors (Navani, Sharma & Sapra, 2012). The main causes of the existence of technical losses in electrical systems are mentioned below:

- Joule effect
- Corona effect
- Harmonic distortion

2.2.2. Non-technical losses of electrical energy

Correspond to the amount of energy that is delivered to final users but is not economically represented in the cash flow of companies in the electricity sector

(Navani, Sharma & Sapra, 2012). This type of losses are mainly associated with errors in the reading and billing processes of the companies, errors in the measurement equipment and with the fraudulent behavior of the users who deliberately manipulate the elements of the system to decrease or evade the payment for the use of the electric power service.

Due that the central axis of this research corresponds to the problems associated with the non-technical losses of electrical energy that are the product of the fraudulent connections of the users, the concepts associated with this problem will be described in a greater degree of detail.

2.2.2.1 Main electrical connection

It is defined as the set of conductors that are responsible for supplying electrical energy to the end user. These are derived from the distribution network and connected to the client's property. There are two types of connections, the aerial type where the conductors are visible and exposed to the air during their entire path, and the underground in which the conductors travel through the interior of ducts to the point of connection (Rios, 2013). Figure 2.2-1 shows an image of the type of conductor used in main electrical connections.

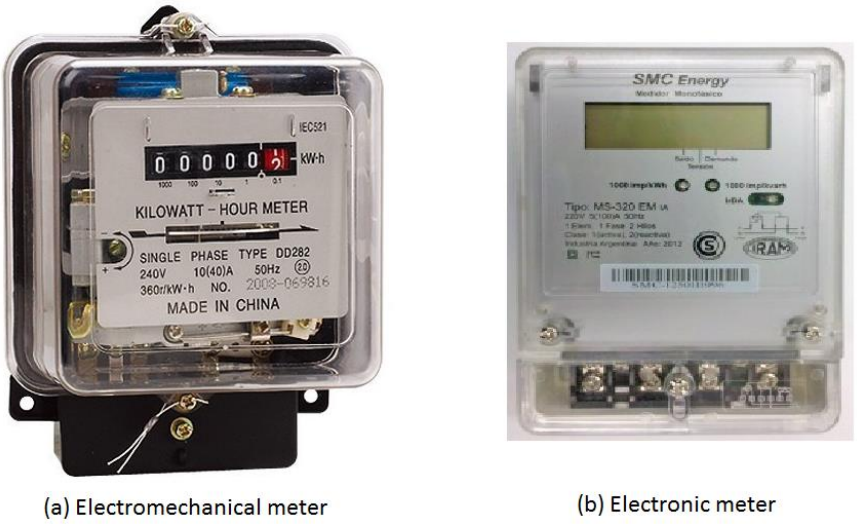
Fig 2.2-1. Cable used in low voltage main connections. (www.nexans.co)

2.2.2.2 Electric meter

It is the device in charge of carrying out the measurement of the customer's electric energy consumption. Previously electromechanical meters were used, which are currently being replaced by electronic meters. There are meters of different types depending on the level of voltage and current of the user. They are

divided mainly in direct, semi-direct and indirect measurement (Rios, 2013). The direct measurement devices are directly connected to the user circuit because they are able to handle the voltage and current levels of the system. The semi-direct measurement devices are able to handle either of the two magnitudes, either voltage or current, while the other exceeds the operational limits of the device. Indirect measurement devices are not capable of directly sensing either of the two magnitudes because both exceed the limits of operation of the equipment. For cases of semi-direct and indirect measurement, it becomes necessary to use transformation devices to be able to bring the variables to manageable values for the measuring equipment.

Figure 2.2-2 shows images of an electromechanical meter and an electronic meter, although there are a great variety of marks and designs, they generally look in a very similar way to those presented below.



(a) Electromechanical meter (b) Electronic meter

Fig 2.2-2. (a) Electromechanical meter, (b) Electronic meter. (www.smc-energy.com)

2.2.2.3 Auxiliary elements of the meter

They are defined as those devices that must be added to the meter in order to be able to perform correctly the measurement in cases where this on its own could not. Within this definition are the potential transformers (PT) that are responsible for reducing the voltage level of the supply circuit to a value that is within the operating limit of the meter, and current transformers (CT) whose function is

similar to that of the potential transformers but associated with the current decrease. These types of elements are used in cases of semi-indirect and indirect measurements.

2.2.2.4 Fraudulent behavior

This concept refers to any effort of the client to manipulate one or more of the parts that compose the electricity supply system (main connection, meter or auxiliary elements), in order to reduce or avoid payment for the use of the Electric power service (Suriyamongkol, 2002).

2.2.2.5 Fraudulent connection

It corresponds to the materialization of the fraudulent behavior of the customer. Is represented in any physical alteration of the electricity supply system either by modifying or breaking any of its parts, as well as adding some external element. Fraudulent connections can be of the craft type or through the application of technical knowledge in electricity, this has led to various types of fraud, making it increasingly difficult to detect them.

Fraudulent connections are classified according to the element of the supply system where they are located. The following figure shows an outline of the typologies of fraud according to the affected element:

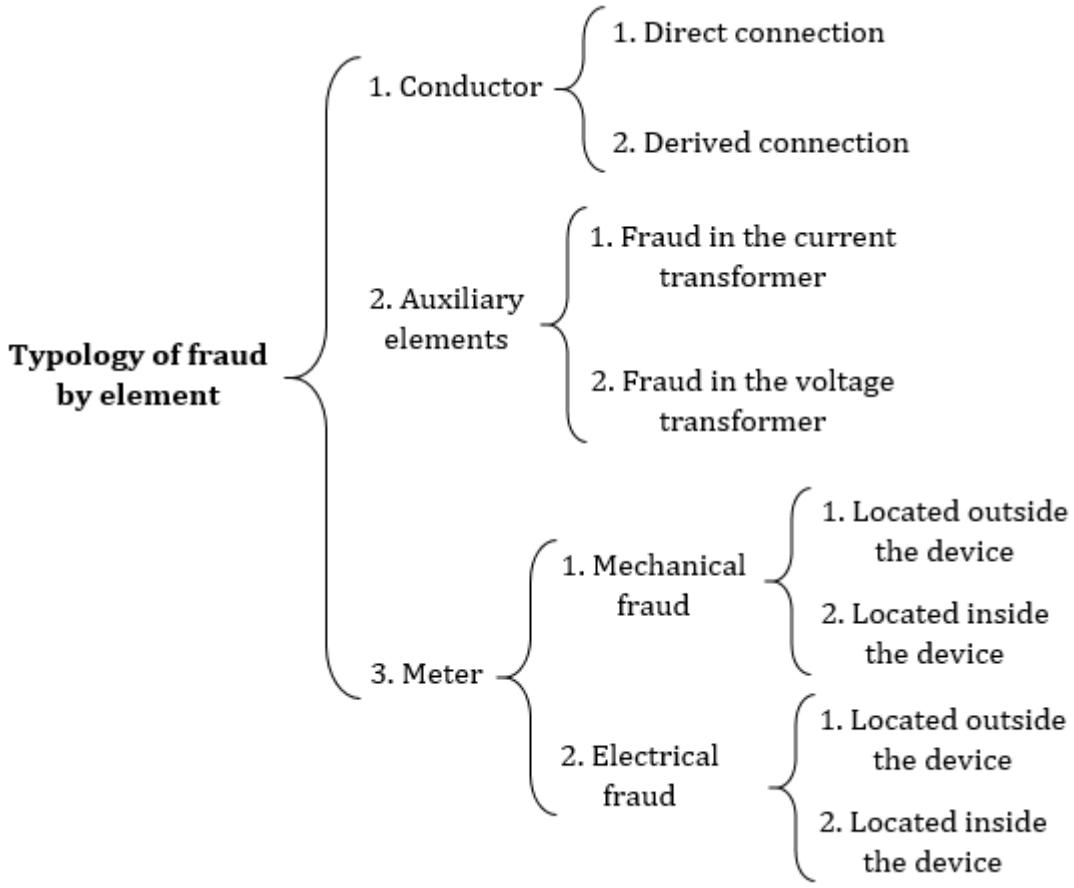


Fig 2.2-3. Scheme of the classification of frauds according to their location.

For each typology of fraud presented in figure 2.2-3, there is a wide variety of modalities in which these can be presented, giving rise to a significantly extensive list and whose detailed description of its elements is not the main objective of this chapter. However, examples of the most common instances of each type of fraudulent connections presented above will be listed.

Frauds in main electrical connection

The main types of fraud in the main connection are summarized below (Rios, 2013):

- Direct connection to the distribution network through the use of a conducting element external to the supply system (hooks, needles, nails, barb wire, etc.).
- Direct connection by means of double shot blade.

- Direct connection through the isolators that support the distribution network.
- Derivation of the connection through the use of some conducting element external to the supply system (hooks, needles, nails, barb wire, etc.).
- Derivation of the connection through clamps.
- Derivation of the connection via transfer system.

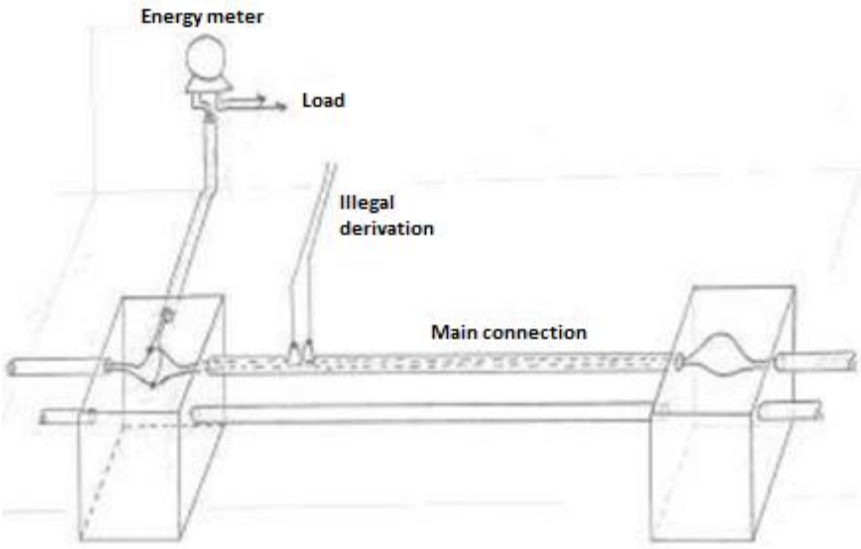


Fig 2.2-4. Fraud by derivation of the underground connection.
(www.afinidadelectrica.com)

Frauds in auxiliary elements

The main types of fraud in the auxiliary elements are summarized below (Rios, 2013):

- Change of transformation ratio of current and potential transformers.
- By-pass connection on one of the coils.
- Open voltage signal (disconnected).
- Rectification of current or voltage signals.
- Auxiliary phase conductor to counter current flows in the transformer.

Frauds in the electric meter

The main types of fraud in the meter are summarized below (Rios, 2013):

- Gear reduction in electromechanical meter.

- Variation of tilt angle of the electromechanical meter.
- Modification of any component of the mechanical system (pinion, gears, disc, shaft, etc.).
- Disconnection of the neutral conductor from the meter.
- Insertion of resistance in series with the measuring circuit.
- By-pass conductors on the meter coils.

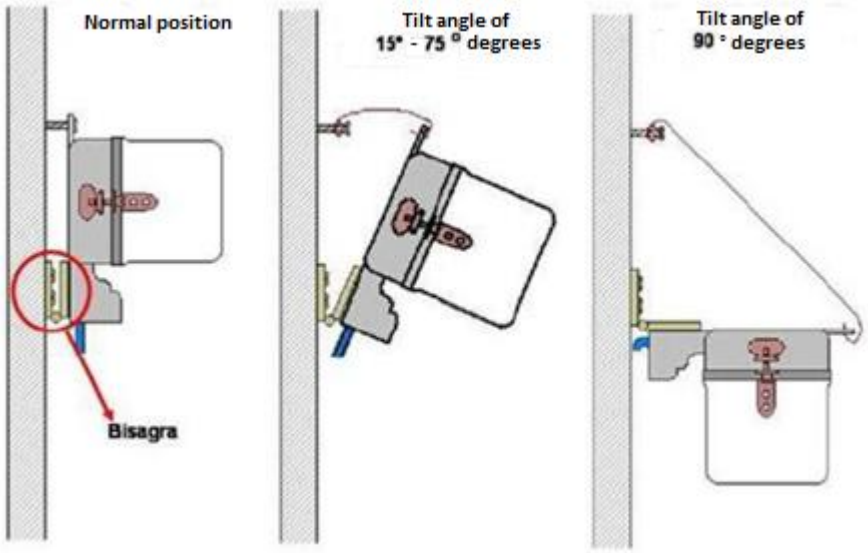


Fig 2.2-5. Fraud by variation of tilt angle of the meter. (www.afinidadelectrica.com)

Concepts associated with the approach of this research to solve the problem under study

2.3. Machine learning, statistical models of time series and computational intelligent techniques

2.3.1. Machine learning

It is defined as machine learning to the internal changes that occur in systems related to tasks associated with artificial intelligence, these changes can affect its structure, code and information, and are the product of an iterative process in which the system tries to refine itself to obtain an adequate output from a stipulated set of inputs, allowing it to improve its future performance in the task for which it is being used (Nilsson, 1998). There are two approaches of machine learning, supervised and unsupervised learning, which are described below.

2.3.1.1 Supervised learning

It corresponds to the machine learning approach in which the objective is to obtain a function that describes in a specific way the data that make up the set of inputs. However, for this particular case, this objective is achieved through a training process based on examples, i.e., for each value of inputs is known its output, so that the system generalizes a function that adjusts in the best possible way the inputs with their respective outputs. The term supervised is used because the system is supplied with the values of the outputs for each value of the inputs during the learning process (Nilsson, 1998).

2.3.1.2 Unsupervised learning

The objective of this type of learning is essentially the same as supervised learning, with the difference that in this case the system is only supplied with the set of values of the inputs and not the values of the outputs. Therefore, the process is quite complex because the obtaining of the function that describes the input data set arises from an intrinsic analysis of the same, in order to find patterns and relationships between them to obtain the possible outputs. The above is quite different from the supervised case because in this case the learning is given by means of examples, thus allowing the function to be obtained by generalizing the values of the outputs to their corresponding inputs (Nilsson, 1998).

Machine learning is used for several tasks, however, for the development of this research three of them were used as it will be described next.

2.3.1.3 Regression

In the field of machine learning, the concept of regression is associated with the use of a model based on one or several computational intelligence techniques to obtain an objective function, that function must be the one that best describes the output of a process based entirely in the relation present in the inputs of the same. That is, an intelligent algorithm that is able to generalize the output function described by the interaction of the input data is used (Dobra, 2005). This concept is the same that is used in statistics, with the difference that the means to reach the solution is not a statistical method but a computational intelligence technique.

Within the regression applications are the curve fitting, as well as the prediction and forecast of processes such as time series.

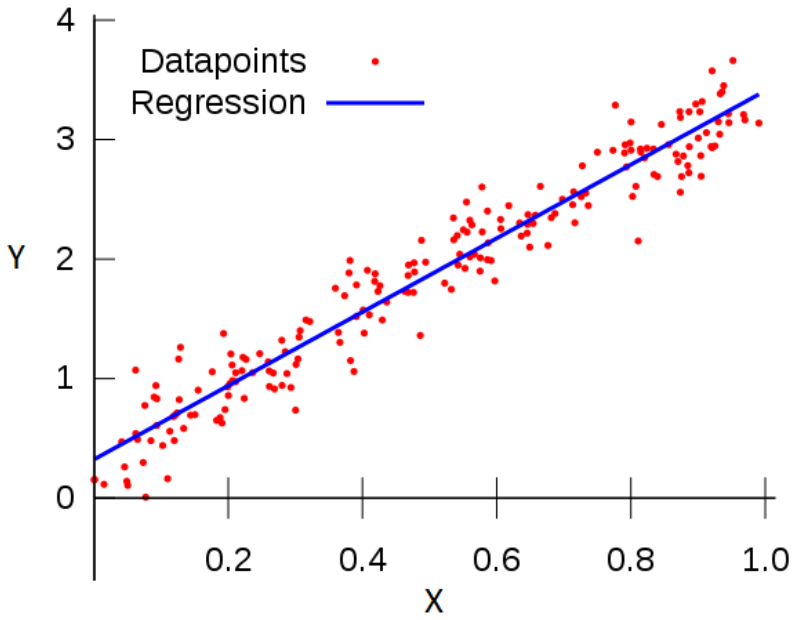


Fig 2.3-1. Illustrated example of a regression model. (www.medium.com)

Figure 2.3-1 shows a process characterized by two input variables X and Y. In addition, the process outputs are known for a set of combinations of the inputs, these outputs are represented with red points in the plane. The blue line represents the function obtained from the regression analysis performed, which is the function that best describes the outputs of the process from the values of each combination X and Y in the input.

2.3.1.4 Classification

It is conceived as an analogous problem to regression, where the main objective is also to find an objective function through the use of computational intelligence techniques. However, the difference is that the function to be found is not the one that makes a better adjustment of the output of the process from the interactions of the inputs, but turns out to be the function that allows a better separation of the different outputs of the process (Dobra, 2005).

In classification problems the outputs of the process are discrete. They correspond to point values that define labels or tags of different classes. Normally a set of input variables that define the membership of each element of the sample

to a class or another is given. The task of the algorithm is to find the function that allows to separate better the different classes that can take the outputs, allowing a correct classification of each element in its corresponding group. Figure 2.3-2 shows an example of a classification function obtained to separate elements belonging to two distinct groups.

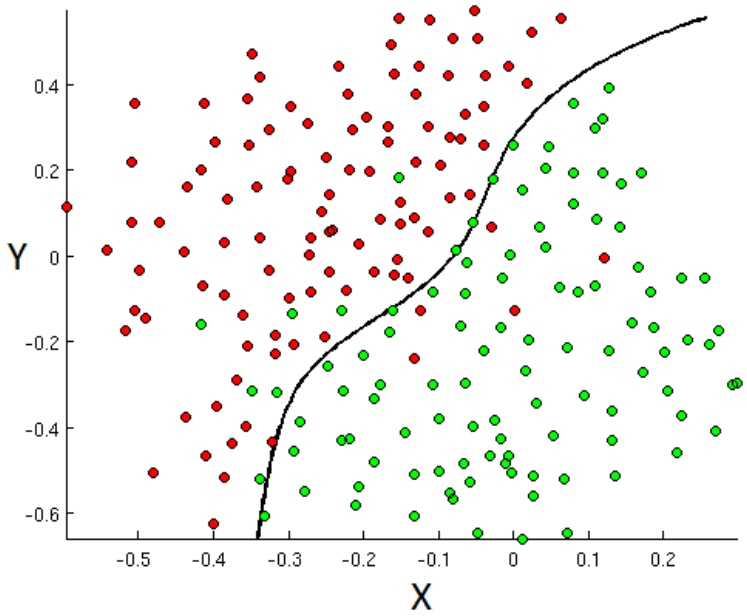


Fig 2.3-2. Illustrated example of a classification model.
(www.openclassroom.stanford.edu)

From the previous figure can be observed that there is a process described by two variables X and Y, which define the membership of each element of the output in one of two groups (green or red). The curve line represents the classification function that allows to separate the points belonging to each group. However, although it does not make a perfect classification, it is the best possible function that can be obtained to classify the set of elements of the sample.

2.3.1.5 Clustering

The tasks described above (regression and classification) belong to the supervised learning approach. In this sense, the intelligent algorithm is able to adjust an output function from a set of values in the inputs and knowing the value of the outputs of the process for each combination of them. Clustering is a task that

belongs to the unsupervised approach, that is, only the values of the inputs and not the outputs are known. In this case, it is desired to perform a grouping of a set of elements from the relationships and patterns present in the input variables that characterize them, without knowing previously to which group each belongs. The task of the intelligent algorithm is to find the function that allows to perform such grouping by identifying similar characteristics present in the set of inputs, such that the elements present in each group are similar to each other and different from those present in the other groups (Nilsson, 1998).

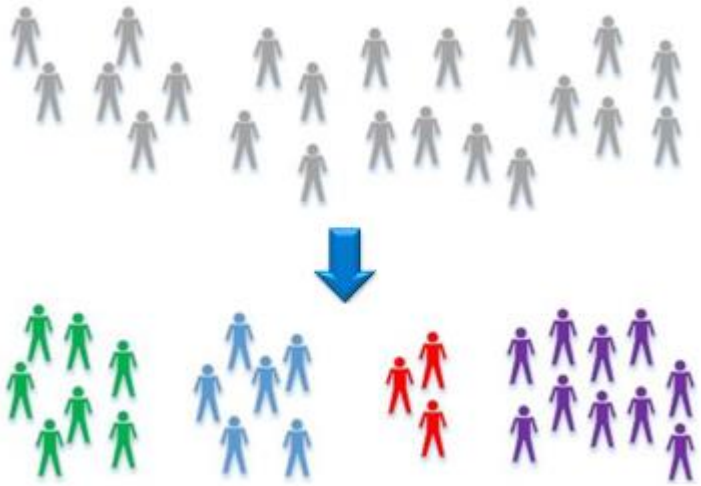


Fig 2.3-3. Illustrated example of the clustering process. (www.insidebigdata.com)

Figure 2.3-3 shows an example of the process of clustering, in which a set of people described by an unknown number of variables is given and the idea is to be able to group them so that the individuals that result in each group are as similar as possible, and at the same time differ significantly from individuals in other groups. Before the execution of the algorithm people are shown gray, this to imply that it is not known to which group they belong. However, after the execution of the same, four different groups are designated with the colors green, blue, red and purple. Its being understood that those that resulted in the same group share significantly similar characteristics in terms of the variables used to describe the set of people.

2.3.2. Statistical modeling of univariate time series

The models used in this research for the analysis of univariate time series are described below.

2.3.2.1 Autoregressive moving average model (ARMA)

It is an explanatory model of time series which results from the combination of an infinite impulse response filter or autoregressive (AR) model, and a finite impulse response filter or moving average (MA) model (Adhikari & Agrawal, 2013), its general expression is presented below:

$$X_t = \varepsilon_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i}. \qquad (2.3\text{-}1)$$

The autoregressive part of the model allows to explain an observation of the series by a linear combination of a certain number of its past observations, the number of observations to be taken into account is given by the value of the constant ***P*** and determines the degree of the autoregressive polynomial of the model. On the other hand, the moving average model is based on the idea that it is also possible to explain an observation of the time series by a linear combination of stochastic impulses. These are also known as innovations and corresponds to error terms from a normal distribution; the number of pulses to be taken into account is given by the constant ***Q***, which determines the degree of the moving average polynomial of the model (Klose, Pircher & Sharma, 2004).

Theoretically, there is no deterministic way to obtain the ideal values of ***P*** and ***Q*** for the creation of the model. However, in practice several ARMA models with different values are usually adjusted to find the one that best fits the time series. In addition, there is a somewhat better grounded approach which is based on the study of the autocorrelation and partial autocorrelation functions of the time series to obtain the degrees of the AR and MA models. However, a heuristic technique known as the Akaike information criterion (AIC) (Hu, 2007) was used in the proposed intelligent system due to it was necessary that the adjustment of the model to be made automatically.

2.3.2.2 Autoregressive integrated moving average model (ARIMA)

This method is essentially an improvement of the model previously treated, it is often the case that many time series describe a non-stationary behavior and in these cases ARMA models do not perform well. Due to this, the existence of the integral term in the expression of the model is proposed, which is nothing more than the application of the finite difference method to the time series to guarantee the stationarity of the same and to obtain a better performance of the adjusted model (Adhikari & Agrawal, 2013; Klose, Pircher & Sharma, 2004).

Otherwise, the ARIMA models are exactly the same as the ARMA models, therefore their main features have already been described previously. The following is the general expression of this type of model:

$$Y_t = -(\Delta^d Y_t - Y_t) + \phi_0 + \sum_{i=1}^{p} \phi_i \Delta^d Y_{t-i} - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \quad (2.3\text{-}2)$$

In general, it is sufficient to use an integral term equal to one (1) to guarantee the stationarity of the ARIMA process. The values of ***P*** and ***Q*** that allow an acceptable fit of series such as the electric consumption whose behavior pattern takes place within a small range of samples (12 in this case) are usually values between 1 and 2, which will be determined with the Akaike information criterion (Hu, 2007).

2.3.3. Computational intelligence techniques

The soft computing techniques used in this research are described below.

2.3.3.1 Genetic algorithm (GA)

This term is used to refer to a set of algorithms whose purpose is to solve an optimization problem by imitating the biological process of natural selection. Genetic algorithms belong to the group of metaheuristics associated with so-called evolutionary computation, which are optimization techniques inspired by the behaviors of nature. In an analogous way to the natural evolutionary process, the genetic algorithm simulates the phases of the same, this through the use of the basic genetic operators such as selection, crossover and mutation. Figure 2.3-4 shows the flowchart of this computational intelligence technique.

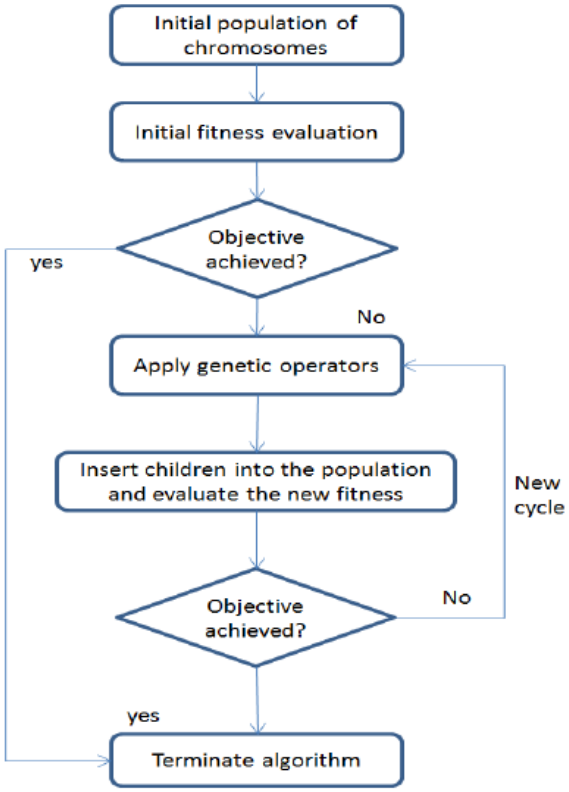


Fig 2.3-4. Flowchart of the Genetic Algorithm. [23]

The process begins with the establishment of the initial population, which for the problem to be optimized corresponds to the initial set of combinations of system inputs, then each of these combinations is evaluated to establish which generates the best solutions. If the desired solution is not found, the genetic operators are applied to introduce modifications in the population, this step is repeated until finding the most suitable solution, which will correspond to the optimum desired. Each of the genetic operators is briefly described below.

Selection

At this stage of the process the best individuals of the current population are selected depending on their results according to the fitness function. This step guarantees the reproduction of individuals with better characteristics, that is, guarantees the evaluation of those combinations of inputs that are closer to the desired optimum. The above is based on the law of survival of the fittest, where superior individuals reproduce to generate children with equally or even better characteristics (Karray & Silva, 2004).

Crossover

Similar to the natural case, crossover consists of the pairing of two chromosomes, these must clearly come from individuals favored from the selection process. This stage is based on the concept of reproduction, where the crossing of individuals with good characteristics should generate children of equal or better characteristics. This allows to introduce modifications into the initial population, allowing a more complete and detailed exploration of the universe of possible solutions of the process (Karray & Silva, 2004).

Mutation

In nature, mutation is the partial modification of a chromosome during crossover, this changes the genetics of the resulting individual in an unpredictable manner, which may be favorable or unfavorable for said individual. In the case of the genetic algorithm, the mutation occurs in a similar way to the one mentioned above, randomly altering the information of one of the child's chromosomes. This is used in this technique to allow the exploration of new solutions, other than selection and crossover, since these allow the exploration of variants of existing solutions (Karray & Silva, 2004).

2.3.3.2 Artificial neural networks (ANN)

This computational intelligence technique is inspired by the structure of the human brain, which is made up of millions of interconnected nerve cells known as neurons. These interconnections form a network that is responsible for the human being to be considered intelligent, given that it provides thinking, reasoning and decision-making skills that enable him to face any kind of situations.

In a similar way, artificial neural networks have a fundamental basic unit known as perceptron, which is nothing more than the mathematical representation of a neuron. In order to obtain a system with human abilities, the perceptrons are interconnected by layers, which together form a network similar to that found in the brain, hence they are called artificial neural networks (Matich, 2001). Figure 2.3-5 shows the basic structure of an ANN, which consists of a layer of input neurons that is responsible for receiving the input variables and distributing them

to the interior of the network, a hidden layer that performs the mathematical operations and an output layer responsible for making the final decision with the results obtained from the hidden layer.

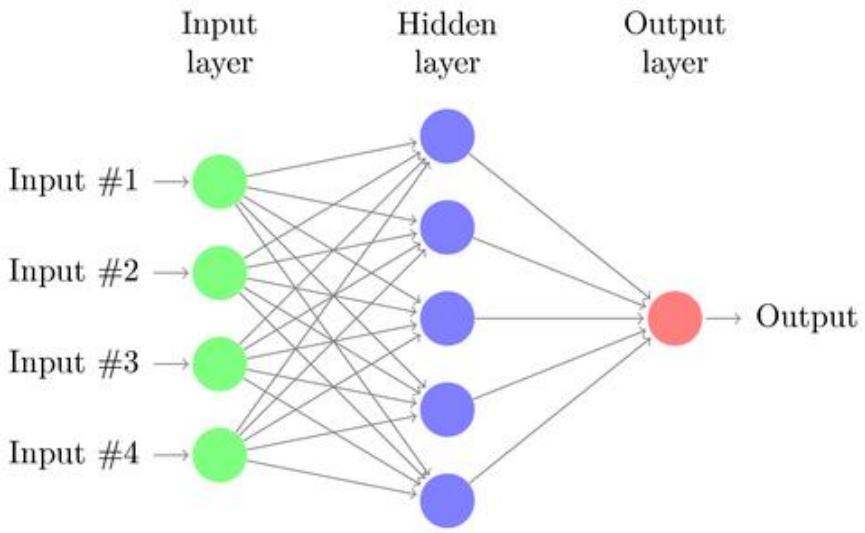


Fig 2.3-5. Basic structure of an ANN. (www.quora.com)

Within the capabilities of artificial neural networks is the ability to learn from experience, generalize from examples, extract patterns and find relationships in data sets. This technique, like all those belonging to machine learning, requires a training process to adjust its internal structure, so that it is possible to obtain the correct response of a process or system to a set of excitation variables in the inputs (Matich, 2001).

2.3.3.3 Self-Organizing Maps (SOM)

It is a specialized neural network topology for solving unsupervised clustering problems. It is based on the fact that the elements of a sample can be characterized by a set of ***N*** variables. The fundamental idea of self-organization maps is to reduce the dimensionality of the problem, that is, to reduce the number of variables necessary to describe each element from any number ***N*** to 2. This allows the mapping in two dimensions of the input data, which is necessary for the execution of the technique.

Once the data are in the two-dimensional space, a number ***G*** of neurons interconnected with each other is randomly distributed in said space, that number corresponds to the number of groups in which it is desired to divide the sample of elements. After this, it starts an iterative process in which it is sought to cover the space demarcated by the dispersion of the data with the network of neurons, the process ends when the network has completely covered the data set. Each neuron represents a group, therefore, the distance of each element is evaluated to all the neurons of the network, placing that element in the group corresponding to the neuron with the shortest distance to it (Van Hulle, 2012). Figure 2.3-6 illustrates the iterative process of training a self-organizing map.

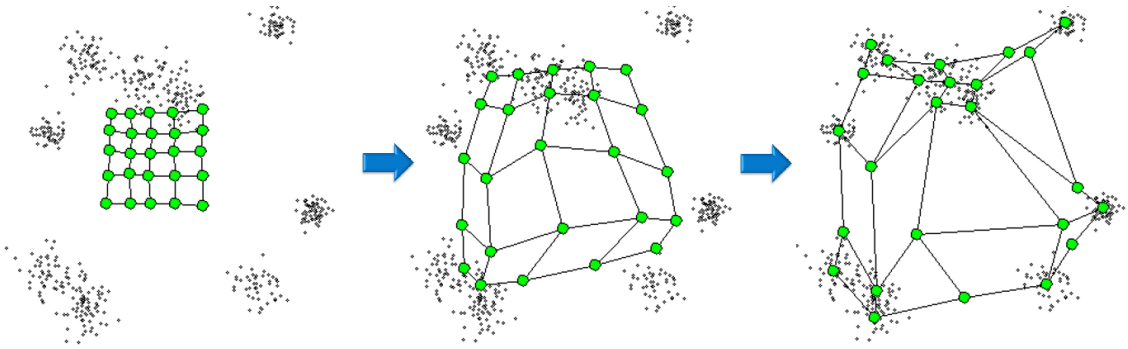


Fig 2.3-6. Illustration of the training process of a SOM. (www. bioinf.fz-borstel.de)

Once the map is trained, the grouping of new data is given again by evaluating all the distances between each neuron of the network and the data to be grouped, the neuron with the shortest distance to the data absorbs it in its respective group. This is illustrated in Figure 2.3-7, which evaluates three new elements that wish to be grouped with the self-organization map presented in the figure.

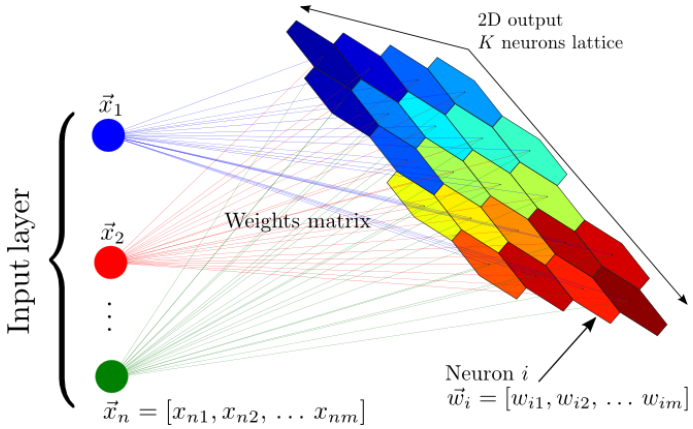


Fig 2.3-7. Grouping of three new elements in an already trained SOM.
(www.matias-ck.com)

2.3.3.4 Support vector machines (SVM)

At this point, the reader might begin to believe that all computational intelligence techniques are directed toward the algorithmic replication of natural processes. This is not true, not all techniques belong to the paradigm of evolutionary computation, such is the case of the technique that will be described below. Support vector machines are a type of soft computing technique whose purpose is to achieve the representation of the problem to be solved by mathematical models of optimization.

It was initially created as a supervised classification algorithm, therefore, its general definition is oriented towards the solution of this type of problems. Nevertheless, the approach of this technique has been modified from the general definition to allow it to solve also other types of problems within the supervised approach such as regression and curve fitting.

Following its general definition, the SVM algorithm represents the input data as points in a space with a dimension corresponding to the number of variables that characterize the input elements. The objective is to find a hyperplane in this space that allows a linear classification of the set of elements that make up the problem. However, as shown in figure 2.3-8, it is not always possible to classify linearly the points on the original space of the data. To achieve this, the SVM always maps the points from the original space to a space of very high dimensionality in which it is possible to perform a linear classification of the elements (see figure 2.3-8). It should be noted that the points closest to the obtained hyperplane are known as support vectors, and are those that give the name to the technique (Andrew Ng, 2015).

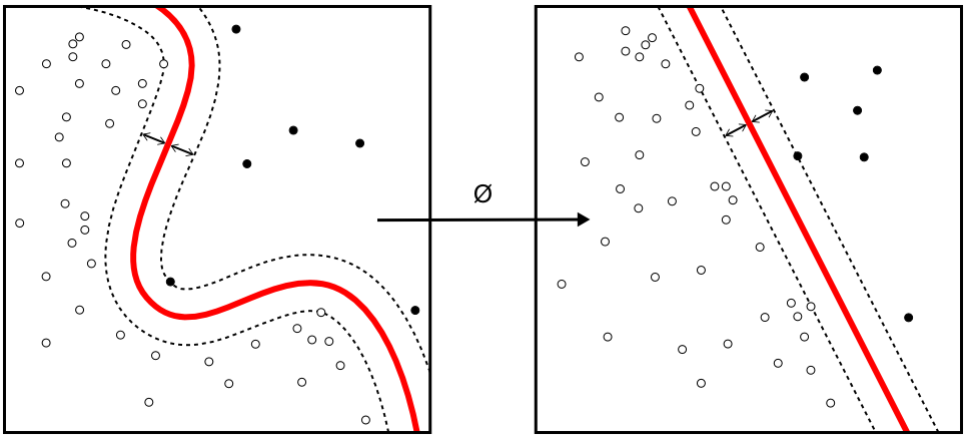


Fig 2.3-8. Mapping of elements during SVM execution (www.wikipedia.com)

Although the mapping of the elements to a space of high dimensionality contributes to solve linearly the problem, this negatively impacts the execution of the algorithm because to perform the computation of the calculations and operations in that space is significantly more expensive. Fortunately, the SVM makes use of an operation known as the Kernel trick, which allows to write the optimization problem that is being solved as a set of dot products, which are defined in high-dimensional space but can be calculated in the original space through special functions called Kernel functions. In this sense, the problem is solved linearly, in a space of high dimensionality, but doing the calculations and operations in the original space to avoid the high computational cost (Andrew Ng, 2015).

2.3.3.5 Random forests

It is a technique derived from that known as decision trees, which adds characteristics of a paradigm known as bagging and introduces some substantial modifications to carry out its tasks in a more efficient way in comparison with the two previously mentioned.

Suppose a set of elements where each one of these can be characterized by a finite set of variables, these variables being common to all elements of the sample. In the field of machine learning, a decision tree is a model consisting of an initial node where it is asked for the value of one of the characteristic variables of the elements, and depending on the response branches to arrive at a new node where it is asked

for the value of any of the other variables to proceed to take another branch. This process of questioning, decision and branching is repeated successively until the path is finished, where there are sufficient reasons to make a decision regarding the problem that gave rise to the implementation of the tree. Figure 2.3-9 shows an example of the typical structure of a decision tree, in which an individual asks if he will be able to hit or miss an event based on the decisions that can be made for each possible scenario.

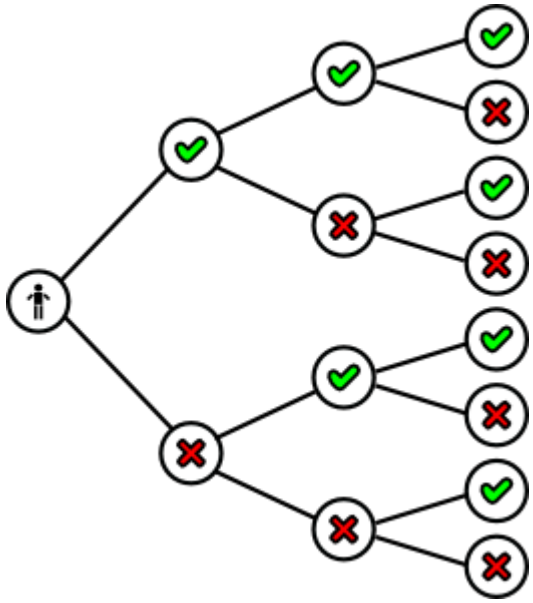


Fig 2.3-9. Typical structure of a decision tree.

In order to be used as a computational intelligence technique, a decision tree must be trained through an iterative process that guarantees the best possible tree, that is, it is capable of reaching the response with the least number of branches. It should be clarified that due to the above, even if it is desired to solve the same problem from the same data set, if these data are partitioned differently for the training and validation of the model, completely different trees will be obtained at the end of the process (Tan, Steinbach & Kumar, 2006).

The above fact is the fundamental concept of a derived approach known as Bagging trees or decision forest, in which a varied set of decision trees are trained independently and randomly to finally average the output of each one and obtain the response of the model. The aim of this approach is to increase the precision of

the model by combining a large number of trees, thus reducing the variance of the system. (Tan, Steinbach & Kumar, 2006). Figure 2.3-10 shows a graphical representation of the decision forest approach, where a total of ***T*** decision trees were used in the model construction.

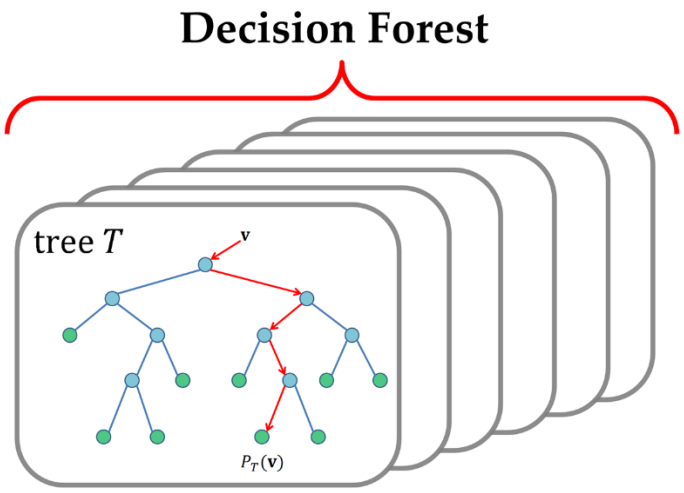


Fig 2.3-10. Graphical representation of the decision forest approach (www.mathworks.com)

Properly defining the technique of random forest, it can be presented as a modification of the bagging technique that has been shown to obtain better performances when compared to each other. The difference between the two is that for the training of each bagging tree all the variables that characterize the elements of the sample are used. On the contrary, for the training of each tree in random forest a fixed random number of variables that is less than the total of the same ones are selected. This reduces the correlation between trees due to the attenuation of the effects of those variables that most influence the output of the process, leading to better results due to the complete exploration of the interaction between these variables (Breiman, 2001). Figure 2.3-11 illustrates the difference between bagging and random forest, presenting an example in which we have ***M*** variables to characterize a set of elements, which are object of study through the implementation of both models. It can be seen that in the bagging all the ***M*** variables are used while in random forest a number ***m*** of them are randomly selected.

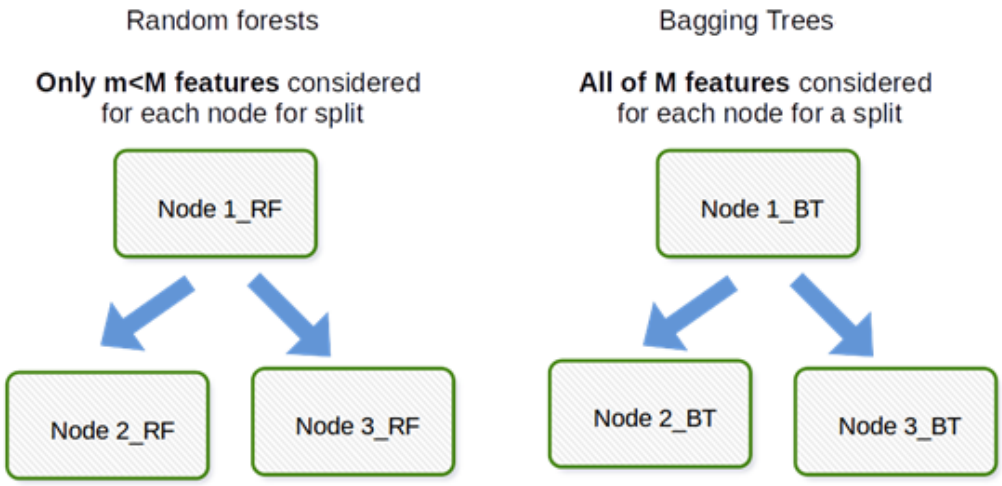


Fig 2.3-11. Random forest and bagging division process at each node (www.quora.com)

Chapter 3

Related Work

This chapter presents an overview of the main works focused on the topics addressed in this dissertation.

3.1. Systems based on computer intelligence techniques for detection of fraudulent users

(Czernichow, Muñoz & Sanz-Bobi, 1998) were among the pioneers in proposing a methodology for detecting fraudulent users through the use of computational intelligence techniques. The model is based on the use of an unsupervised neural network to recognize and group the different consumption patterns of the users, in order to identify those that represent fraudulent behaviors.

The proposed methodology is described by the following steps:

1. **Input variable:** The variable determined by the authors was the record of historical consumption (consumption profiles) of each customer in a period of twelve months.
2. **Information filtering:** There were processes of elimination of erroneous information, homogenization of users and normalization of consumption profiles. The homogenization allowed to execute the system for a specific type of users, while the normalization was oriented to capture the consumption behavior of each profile without the affectation produced by the scale.
3. **Model adjustment:** The training of a probabilistic neural network with radial base function was executed in this step. This allowed to extract

common patterns between groups of profiles. The above with the purpose of establishing the groups in which the set of users was divided.

4. **Analysis of the model:** Once the groups were identified, the characteristic patterns of each user were evaluated to determine which group they belonged to. The detection of fraudulent users was carried out in the following two ways, selecting those users who were located in groups whose behavior patterns were known to be anomalous and selecting those that did not belong to any of the established groups.
5. **Evaluation of the model:** Although the authors do not reach a specific result, nor do they perform a validation of the model, a coefficient of normality is obtained for each user. A distribution curve of the coefficients of normality obtained is presented, where it is possible to see that those that are far from the conglomerate can be classified as fraudulent.

(Jeffrey, Jiang, Lachsz & Tagaris, 2002) proposed a hybrid model between a technique of time series analysis known as wavelet decomposition and an artificial neural network, in order to perform a classification of users into two groups, normal and fraudulent. The model receives the consumption profiles of each client, to which wavelet decomposition is applied to obtain its representative parameters, these parameters are entered as descriptors to the neural network in charge of performing the classification.

The methodology of the authors is described by the following set of steps:

1. **Input variable:** The only variable taken into account by the authors is the historical record of consumption (consumption profiles) of each customer during a period of one year, with data produced by measurements taken at fifteen minutes intervals.
2. **Pre-processing of the information:** This step is carried out to clean the information and eliminate the erroneous data, so as to guarantee the quality of the information that enters the system.
3. **Extraction of characteristics:** To each profile of consumption is applied the wavelet decomposition, this technique allows to realize a series transformation in the time domain to a series in the wavelet domain, which

allows a simplified representation taking into account only the characteristic parameters of the original series.

4. **Classification:** The classification phase was carried out through an artificial neural network, which received as descriptors of each client the parameters extracted from the wavelet decomposition. The output was established in binary representation, allowing each user to be described as fraudulent or non-fraudulent.
5. **Results:** The authors report that after the simulations, the results obtained for the training phase of the network are maintained at an average of 78%, and those obtained in the test phase are on average 70%. This places the proposed methodology as a good tool for detecting fraudulent users.

(Cabral, Gontijo, Pinto & Reis, 2004) developed a model for the detection of fraudulent users based on the Knowledge Discovery in Databases (KDD) methodology, which proposes a set of steps that must be followed for the correct solution of a problem based on a dataset (information system). The selected computational intelligence technique was rough sets, whereby a set of data described by $\boldsymbol{M}$ variables can be reduced to a simplified dataset with $\boldsymbol{m}$ variables ($\boldsymbol{m} < \boldsymbol{M}$), which contains the same information and allows to reach the solution much simpler than the original set.

The KDD methodology followed by the authors is described below:

1. **Attribute selection:** Ten variables were selected to characterize each client of the exploration space of the proposed system. The authors claim that these variables were obtained from the experience of people working in commercialization companies in the electricity sector.
2. **Pre-processing of the information:** A delimitation of the universe of customers was made, from which it was decided to work only with those that have been visited by inspection crews and the result of this inspection is known. The time window to be analyzed was limited to a total of twelve periods (one year) per customer.
3. **Transformation of the information:** This stage is used to obtain new variables that are result of the combination or the statistical-mathematical

operation over the initials. An example of this is to create an average consumption variable from historical consumption data.

4. **Data Mining:** At this stage, the computer intelligence technique is applied, which allows users with fraudulent connections to be detected. As mentioned before, the technique used was rough sets, which allows to create a set of rules whose purpose is to facilitate the desired classification of users from a simplified dataset, which is obtained by a reduction of the original set.
5. **Evaluation of the model and results:** The execution of the intelligent algorithm allowed to obtain a percentage of successes in the detection of fraudulent clients of 20%, this was considered sufficient by the authors because the company with which they worked together had a percentage of 10% before the implementation of the system. It was concluded that the results obtained were due to the unbalance of the training database, which had 90% cases of normal clients and only 10% cases of fraudulent clients.

(Hashim, Hussien, Mohamad, Pok & Yak, 2007) proposed a methodology for the implementation of an intelligent system based on a hybrid between a genetic algorithm (GA) and support vector machines (SVM). The technique used for the classification of fraudulent and non-fraudulent users was SVM. However, to obtain the best classification it was necessary to find the optimal values of the parameters of this technique. To solve this, the genetic algorithm was used, which was responsible for executing the SVM iteratively while exploring the different combinations of the parameters of this one until finding the best solution.

The methodology used by the authors is described below:

1. **Input variables:** Only the use of the historical consumption record (consumption profile) of each customer is described, the use of other variables is mentioned, although they are not properly described.
2. **Data conditioning:** Due to the experience in the use of SVM, the exploration limits of the GA are limited so that the search is performed on a set of specific values of the parameters, with which it is known that it is possible to obtain good results in classification problems.

3. **Execution of the intelligent model:** The GA and SVM hybrid model was executed taking into account the following characteristics, eighty-five months of history were used for each consumption profile, the dataset consisted of one hundred and ninety users, it was used 90% for training and the remaining 10% for testing.
4. **Results:** The authors obtained a performance of 94% in the percentage of hits in the classification of users. This may be due to the too small amount of data that the system was running.

(Cabral, Martins, Pinto. J & Pinto. A, 2008) followed the Knowledge Discovery in Databases (KDD) methodology for the proposal of an intelligent system capable of detecting fraudulent connections in users belonging to medium and high voltage. It should be noted that for this type of clients the conditions made the process difficult, this due to the small group of users used for the study, together with the difficulty of finding fraud due to the technification that must be had to connect illegally at these voltage levels. The computational intelligence technique used was self-organization maps (SOM), in order to characterize the consumption profile of each of the users under study. The comparison of the current behavior of a user with the characterization of his historical profile product of SOM allows to detect the existence of anomalous behavior of this one.

The KDD methodology followed by the authors is described below:

1. **Selection of attributes:** The variables used for the characterization of the users were the customer identification number, the daily consumption record of the last year, maximum energy demand, tariff, economic activity and geographical location.
2. **Pre-processing of the information:** It was done a filtering of the data to eliminate the consumptions of the Saturdays, Sundays and holidays; this because for the type of users under study (industrial), these days do not represent their pattern of habitual consumption.
3. **Transformation of the information:** The consumption profile of each client was divided into segments corresponding to a week, so that the SOM was able to characterize the typical weekly behavior of the user.

4. **Data mining:** The SOM was executed for each user of the system, which made a grouping of the series of weekly consumption according to consumption behavior in the same. If it was only possible to obtain a single group it was considered that the user did not change his behavior pattern, which categorized him as a normal user, otherwise, obtaining more than one group suggested a change in the behavior of that week.
5. **Evaluation of the model and results:** Controlled simulations of different scenarios were performed, of which the authors conclude that the system has an effectiveness rate of 85%. The system error is explained by the fact that there are users who present changes in their weekly consumption behavior without being associated with frauds, which are detected as anomalous by the model.

(Fabris, Margoto & Varejão, 2009) presented a methodology based on a hybrid algorithm between supervised and non-supervised approaches. First they solved a problem of unsupervised grouping by a technique known as hierarchical clustering, whereby they divided the set of users into groups according to their consumption behaviors. Next, a supervised classification was performed within each of the clusters by comparing Euclidean distances, thus distinguishing between fraudulent and non-fraudulent users.

The methodology proposed by the authors can be described by the following set of steps:

1. **Input variable:** The variable taken into account was the consumption profile of each user for a one-year window.
2. **Conditioning:** Normalizing of consumption profiles with respect to the maximum of each one was done, in order to obtain series with values between zero and one, from which it is possible to extract only the behavior (pattern) of consumption.
3. **Clustering:** The technique of hierarchical clustering was performed using as a similarity index the distance between curves with the Dynamic Time Warp (DTW) method, quite common in the unsupervised analysis of time

series. This step allows to divide the sample of users into groups according to their consumption behaviors.

4. **Classification:** Within each cluster, consumption series were denormalized to recover their original magnitudes. Next, a classifier based on Euclidean distances was executed to identify fraudulent and non-fraudulent users within each cluster.
5. **Evaluation of the model:** The proposed system is compared with approaches of other authors, of which a table of performances is presented. The model shows a good performance but not the best of those presented in the table. The comparison index used was the F-Measure.

(Figueiredo, Muniz, Tanscheit & Vellasco, 2009) developed a methodology that was based on the implementation of an intelligent system divided into two modules, the first is a filter composed by a committee of neural networks and the second corresponds to a neuro-fuzzy classifier. The idea of the filter is to correct the input data, so that the degree of certainty in the existence of fraud or not in a user before the classification is superior to that obtained from the initial data. The corrected set of data is entered into the classifier, which uses a neural network to adjust the fuzzy rules to decide whether a user is fraudulent or non-fraudulent.

The methodology proposed by the authors can be described with the following steps:

1. **Input variables:** Eighteen variables to characterize the clients were used. These can be divided into attributes of identification, technical aspects, consumption behavior and climatic variables.
2. **Conditioning:** The variables were normalized and atypical, redundant and incomplete data were eliminated.
3. **Filtering module:** Composed of five independent neural networks, the output of the module corresponds to the consensus between the individual outputs of each network. Although from the original data was known whether a customer was fraudulent or not, this information is subject to a large number of possible errors. Therefore, the function of the filter is to clean the original data through the neural network committee, obtaining as

an output the existence or not of fraud but with a higher degree of certainty compared to that obtained from the initial data.

4. **Classifier:** Receiving the corrected data, this module used a neural network to intelligently adjust the rules of a classifier based on fuzzy logic. The creation of these fuzzy rules is usually the task of the system operator based on his experience. However, it is possible to automate the creation of such rules through the use of a neural network. This improves the performance of the module due to the elimination of human subjectivity, creating classification rules based on the analysis of patterns performed by the network.
5. **Results:** The authors stated that before the application of the intelligent system, the commercialization company reported a 32.8% of effectiveness, after the execution of the system it was possible to obtain a fraud detection performance of 51.2%.

(Nizar & Dong, 2009) conducted a study to compare the performance of two computational intelligence techniques in the detection of fraudulent electricity users. The compared techniques were support vector machines (SVM) and extreme machine learning (EML), the latter being one of the most recent and promising approaches of machine learning theory.

The methodology of the authors is described by the following set of steps:

1. **Input variable:** The variable taken into account was the consumption profile of each customer.
2. **Pre-processing of the information:** Elimination processes of erroneous and incomplete data were carried out, as well as normalization of the information.
3. **Identification of profiles:** It is mentioned the use of a clustering algorithm to group the consumption profiles according to their similarity, however, the technique used is not explained in detail.
4. **Classification:** The techniques to be compared (SVM and EML) are applied for each group of profiles obtained from the previous step.

5. **Evaluation of results:** A set of results tables were presented, of which the authors comment that as expected, the classifier based on EML performs considerably better than the one based on SVM.

(Ahmed, Koh, Nagi. F, Nagi. J, Tiong & Yap, 2010) proposed a methodology for detecting fraudulent users based on correlation models. The implemented system is called intelligent by the authors, although it does not use any computational intelligence technique. This is sustained because it is capable of performing automatically and with better performance a process that for a human would be difficult to execute. The model is able to classify fraudulent behaviors based on successive calculations of correlations. These comparisons correlate the daily consumption profile of the client with a set of series that represent anomalous behaviors, which have already been previously characterized by the expert. Obtaining high values in one or more of the computed correlations suggests the existence of a fraudulent connection.

The methodology of the authors is described by the following set of steps:

1. **Input variable:** The daily consumption profile of each client measured in thirty minute intervals was used.
2. **Conditioning:** Wrong and incomplete data were removed.
3. **Execution of the model:** The correlations between the client consumption series and the respective series of anomalous behaviors available are computed. The results are subsequently categorized to classify according to the obtained correlations, which correspond to fraudulent behaviors.
4. **Results:** The authors affirm that after the execution of the system, a performance of 55% was obtained in the detection of fraudulent customers, significantly improving the obtained by the commercialization company, which ranged between 20% and 30%.

(Carmona, Nunes, Saavedra & Silva, 2011) followed the Knowledge Discovery in Databases (KDD) methodology for the implementation of a two-stage intelligent fraud detection system. The first is a fuzzy clustering for the grouping of similar consumption profiles, while the second corresponds to a classifier using a fuzzy membership matrix for the detection of irregular clients.

The KDD methodology followed by the authors is described by the following set of steps:

1. **Selection of attributes:** The average consumption of the last six months, the maximum consumption of the last six months, standard deviation in the last six months, average consumption of the last six months of the locality where the customer is located and an indicator of whether it has been visited for inspection were used.
2. **Pre-processing of information:** Users with more than one inspection visit were eliminated, customers with consumption profiles that exceeded eighteen months were selected and those with incomplete data were eliminated.
3. **Transformation of information:** From step 1 it can be inferred that the transformations were made during the selection of the variables, due to the fact that the average consumption, the maximum consumption and the standard deviation are obtained from operations on the consumption profile.
4. **Data mining:** The fuzzy clustering algorithm was implemented based on a combination of the clustering technique C-Means with fuzzy logic to generate the clustering rules. The above to divide the users into groups according to the similarity of their consumption profiles. Then the second stage is executed, which corresponds to a classifier based on fuzzy logic to obtain the fuzzy membership matrix of each of the clusters resulting from step 1. The Euclidean distance is used as a metric to compare each client with the fuzzy membership matrix, from which it is possible to establish whether a user is fraudulent or not.
5. **Evaluation of the model and results:** Different scenarios were simulated, from which it was possible to establish that the system has on average a 74.5% effectiveness in the detection of fraudulent users.

(Biscarri. F, Biscarri. J, Guerrero, León, Millán & Monedero, 2011) developed a substantial modification of the work presented in 2010 [6]. In this proposal, a methodology was designed for the detection of any type of non-technical loss of

electrical energy, including those products of fraud and also those that occur due to errors in billing, reading or damage of meters. A hybrid system was implemented between statistical techniques and computational intelligence techniques. Fraud detection was carried out by calculating the Pearson coefficient, the classification of other types of non-technical losses was performed using decision trees.

The proposed methodology is described below:

1. **Input variables:** The consumption profile of the last two years, geographic location, as well as the information of readings and billings within the above time window were used.
2. **Conditioning:** Reference is made to the filling of empty data in consumption profiles and filtering to eliminate users with erroneous information.
3. **Fraud detection:** The Pearson coefficient was calculated on the profile of each user to detect drastic decreases in consumption, which are associated with the incidence of fraudulent behavior.
4. **Detection of other types of non-technical losses:** Once obtained the cases of fraudulent users, the rest of users with anomalous behaviors according to the results of the previous stage were used for the execution of a decision tree. The idea behind the implementation of the tree is to be able to distinguish which of the other types of non-technical losses correspond to such remaining cases.
5. **Evaluation of the model:** A percentage of successes in detecting frauds of 38% was obtained. However, the authors conclude the document in a satisfactory way. This was due to the fact that the percentage of the commercialization company was 15% before the implementation of the system.

(Alberto, Costa, Eler, Maduro & Portela, 2013) followed the Knowledge Discovery in Databases methodology to propose a fraud detection model based on artificial neural networks. The classification was performed using a network with

multilayer perceptron topology and trained with the backpropagation algorithm, which has been shown to perform well in pattern recognition problems.

The KDD methodology followed by the authors is described by the following set of steps:

1. **Attribute selection:** Thirteen attributes were used to characterize the users under study. These can be divided into identification variables, technical aspects and consumer behavior.
2. **Pre-processing of information:** Duplicate and incorrect data were deleted.
3. **Transformation of the information:** Normalizations of some of the selected variables were made in this step.
4. **Data mining:** The neural network for supervised classification of users into fraudulent and non-fraudulent was executed.
5. **Validation of the model:** The method chosen was the K-fold cross-validation, which iteratively partitions the input dataset to successively execute the algorithm at each iteration. The performance of the system is the average of the performances obtained in each iteration.
6. **Results:** The authors report a performance percentage above 50% in the detection of users with fraudulent connections.

3.2. Final Remarks

In order to solve a problem from the machine learning approach, it must be kept in mind that in this field the most important thing is the data. Computational intelligence techniques are machine learning algorithms whose strength lies in their ability to process large volumes of data, that is, to examine and analyze them to generalize behaviors and find patterns that make it possible to convert such data into useful information. The advantage of these techniques is that they can use the data to perform tasks that humans are also capable of doing. However, they far exceed the amount of information that any human is capable of processing and in considerably shorter times. In spite of the advantages that these techniques offer, in order to obtain satisfactory results they must be guaranteed an adequate

quantity and quality of data; this becomes the initial step for a correct solution of a problem from the intelligent algorithms viewpoint.

By reviewing the works presented during the chapter, it is possible to conclude that for the solution of the problem of non-technical losses due to fraudulent connections, the authors did not focus on designing an intelligent system but on proposing a methodology to increase the rate of detection of users with such connections. This is evidenced by observing that each proposal is presented through the development of a set of steps, with which it is intended that by following them the process of fraud detection can be improved. It is clear then that the use of an intelligent algorithm is a step further within the adopted approach and not the ultimate end of the process.

In general, the set of steps taken by the authors to contribute to the improvement of the fraud detection process, is similar to any of the standard models for solving a problem from the machine learning perspective. So much that some of them followed one of the most common models for this purpose, which is known as Knowledge Discovery in Databases (Fayyad, Piatetsky & Smyth, 1996) and was already explained briefly during the chapter. Usually, the common steps for solving a problem using machine learning techniques are:

1. **Selection of variables:** In this step the variables by which the elements of the input dataset will be characterized are chosen. They should be those that provide useful information for solving the problem.
2. **Data Conditioning:** This step represents what is mentioned in the opening paragraph of this section. At this point it is guaranteed that the quantity and the quality of the data are optimal for the execution of the computational intelligence technique.
3. **Execution of the intelligent algorithm:** It is necessary to select the computational intelligence technique that is able to solve the type of problem that is being faced. The selection criteria depends entirely on how the solution have been focused, the type of problem, and the designer's experience. Once selected, the technique is executed and the results are recorded.

4. **Analysis of results and validation:** In this step the results obtained are evaluated, in case they are satisfactory the model is validated, if they are not, more tests must be performed. This is an iterative process in which the parameters of the selected technique are configured and executed until obtain results that are valid and allow to solve the problem in question.

The review of all the works related to the topic of interest of this research allows to clearly identify the contribution of this work to the state of the art. The result of this proposal is a methodology belonging to the machine learning approach, which seeks to improve the process of detection of fraudulent users by implementing an intelligent system divided into three stages. Unlike the other proposals, the stages that compose the system of this proposal are not in cascade but the first two are executed independently, and the outputs of these are used as additional inputs to the third, which also receives input variables selected. The first stage is a hybrid cluster between a genetic algorithm and a self-organization map for the intelligent grouping of consumption profiles according to the shape of the curves, that is, their behavior. The second is a consumption predictor based on an ARIMA model corrected by a neural network. The final stage is a classifier based on random forests, whose function is to detect fraudulent users.

Besides the configuration of the system stages, another novelty of this proposal is the inclusion of an optimized clustering (stage 1), which is responsible for finding the best possible grouping according to the similarity between consumption behaviors. Is also novel the implementation of a consumption predictor (stage 2), which is responsible for obtaining the deviation between the predicted value and the real consumption as a measure of possible fraudulent behavior of the customer. Finally, the latest novelty of the proposed approach is the use of a novel classification technique (stage 3) that has not been used for the detection of fraud. This technique is random forest, which has demonstrated good performance in supervised classification problems.

PART II

PROPOSED APPROACH

Chapter 4

Intelligent approach to managing non-technical losses associated with fraudulent connections

This chapter presents the intelligent approach to non-technical losses management proposed in this dissertation applied to the detection of fraudulent users. The main definitions, general considerations, proposed methodology, variable selection and the algorithms for non-supervised grouping of consumption profiles, modeling of consumption profiles and prediction of the next month consumption, as well as fraudulent user detection in this work are presented in this chapter.

4.1. Problem Statement

At this point, it is already clear that the central axis of this research is the problem associated with non-technical losses of electrical energy that are product of fraud by the end users. Such situation affects the operating costs of the electrical system. This is supported by the fact that the commercialization companies must buy more energy than they actually sell, which directly implies that more energy must be generated and therefore more resources must be used to obtain it. Resulting in a chain that affects negatively from the economic to the environmental scope and therefore deserves more attention.

The fraud losses have always existed, however, it has been shown that in Colombia these have been established at a considerably high value in recent years. For that reason, the commercialization companies have developed action plans to recover this lost energy. This represents another additional cost to the operation of

the electrical system, since large amounts of money must be invested in the development of these energy recovery campaigns. In spite of the above, there is no guarantee that the result will be successful because they are based on superficial analyzes of the electrical consumptions. These analyzes allow to schedule the sending of crews for the inspection of the clients in areas where there are indicators of energy losses. In addition, it should be added that fraud methods have become more sophisticated and that not only craft methods but also based on technical knowledge methods are used. In this sense, it's more difficult to detect them.

From all of the above, it is possible to size the magnitude of the problem generated by electric frauds, as well as its repercussions not only in the electric sector but also in the environment. Therefore, it was necessary to handle this problem in a more rigorous way, which prompted the development of proposals from the approach of intelligent algorithms such as those reviewed in the state of the art.

Following the mentioned approach, this research has been addressed to propose a novel methodology for the reduction of non-technical losses that are product of user fraud. The main research component corresponds to the design and implementation of the intelligent system for detection of fraudulent users, which has been oriented from a different perspective to all those previously reviewed. The result is a multi-stage system, with the main stage being an intelligent classifier aimed at detecting users with fraudulent connections. The contribution of this proposal consists in the use of two intelligent stages whose outputs allow to characterize in a more detailed way the consumption behavior of the customers. These outputs are used as reinforcement inputs, which accompany the set of original selected variables as inputs in the classifier.

4.2. Proposed intelligent system for managing non-technical losses associated with user fraud

First it is necessary to define what is referred to by the term "management of non-technical losses", which is part of the title of this research and has been used

repeatedly throughout the document. The word "management" is used in this context to cover the processes of study, analysis and detection of users with fraudulent behavior; which represent most of the non-technical losses of the country's distribution systems and are the focus of this research.

From the review of related works, it was possible to conclude that in order to contribute to the solution of this problem, the authors focus on proposing methodologies and not only on the use of algorithms for this purpose. This is justified because the methodologies involve a set of steps that includes selection of variables, data conditioning and system design; with which it is expected to make a correct characterization of the problem and, therefore, obtain better results. On the other hand, while the direct application of algorithms can lead to satisfactory results, it does not represent an approach or a way of solving the problem and will become obsolete with the emergence of more advanced algorithms.

It is clear then that what is sought as a product of this research is the development of a methodology that allows to improve the process of detection of users with fraudulent connections. In all the reviewed works (see Chapter 3) the implemented methodologies have the structure shown in Figure 4.2-1. In the initial phase, the characterization variables of the problem are selected, entered and conditioned. In the process phase, the computational intelligence techniques are executed in a ***sequential*** manner, and in the results phase it is determined which users are committing fraud according to the output of the intelligent algorithms. The highlighted term refers that in those works where several techniques are used, these are not independent of each other but are executed in cascade.

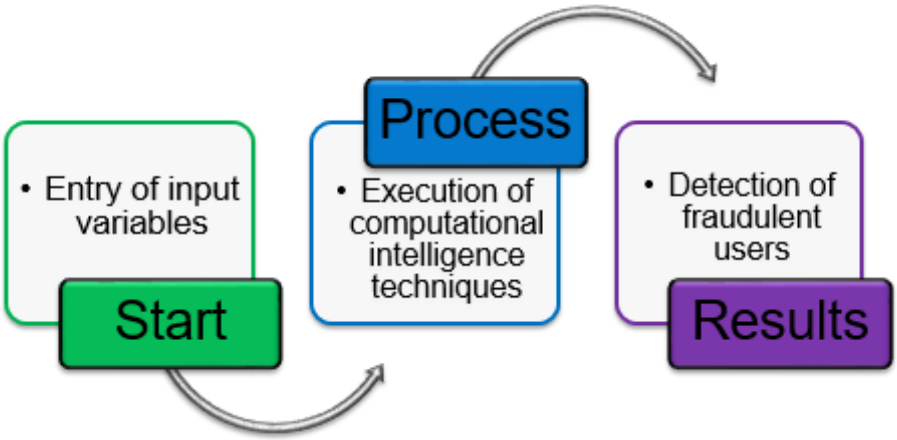


Fig 4.2-1. Structure of the methodologies reviewed in Chapter 3.

The proposed methodology has within its phases the design of an intelligent system composed of three stages (see Chapter 1, Section 1.2.2), of which the first two stages are independent of each other and their results complement the total set of inputs of the system, which are used in stage three. The above is presented as the main reason to make a modification to the methodological structure proposed in Figure 4.2-1. Such modification allows to obtain the structure of the methodology developed in this research and which is presented in the following figure.

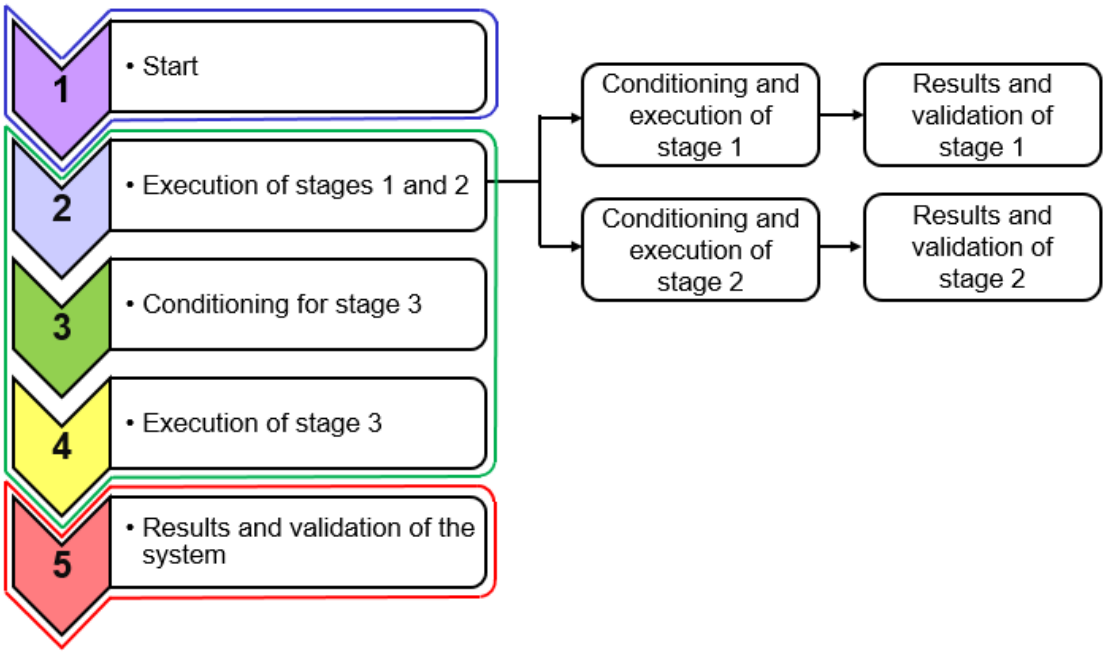


Fig 4.2-2. Structure of the proposed methodology.

From Figure 4.2-2 it can be observed that the methodology proposed in this research has five steps. However, in an attempt to relate it to those representing the standard model of problem solving from the machine learning approach (see Chapter 3, Section 3.2), these steps have been divided into three groups. The first group is marked with the ***blue*** line and contains step one (1), which corresponds to the start phase. The second group is marked with the ***green*** line and contains steps two through four (2-4), which correspond to data conditioning and execution of the intelligent system. The third group is marked with the ***red*** line and contains step five (5), which corresponds to the results phase. Next, the three groups mentioned and the methodological steps that make them will be described in detail.

4.2.1. Group 1: Start

This group contains the methodological step of start (step 1). In this step, the selection of the variables that allow to characterize the problem is made, that is, to define the variables with which the consumption behavior of the users of electrical energy can be analyzed. So that from these it is possible to classify users as fraudulent and non-fraudulent.

It should be kept in mind that this step is the most important of the process, since only a suitable characterization of the problem can obtain proper results. Therefore, it is necessary to select those variables that provide the greatest amount of information about the consumption behavior of customers. The above in order to allow computer intelligence techniques to make generalizations and recognize patterns of behavior, which lead to the correct identification of features that differentiate fraudulent users from non-fraudulent ones.

Given the importance of this step, it was decided to consult the knowledge of experts in the area of control of energy losses. Whose experience contributes to achieve the correct characterization desired. Working together with these experts, it was learned which are the variables that commercialization companies register for the characterization of users. In addition, it was possible to define which of

these variables would allow a detailed analysis of consumption behavior. It is worth mentioning that the information of these variables must exist and be complete for all users of the system. Thus, arriving at the definitive selection of the characterization variables of the problem, which are presented in the Table 4.2-1.

Type of variable	Name of the variable	Property	Measurement
User identification	User identification number (NIS)	Static	Fixed
	User tariff	Static	Fixed
	User measurement equipment	Static	Fixed
	Geographical location of the user	Static	Fixed
Characterization of consumption behavior	Consumption profile	Dynamic	Monthly
	Meter reading	Dynamic	Monthly
	Number of complaints made	Dynamic	Monthly
	Number of overdue bills	Dynamic	Monthly
	Customer fraud history	Dynamic	Monthly
	Reading and billing anomalies	Dynamic	Monthly
Classification of user type	Balance of macrometers	Dynamic	Monthly
	Moorage of macrometers	Dynamic	Monthly
	Results of campaigns and inspections	Dynamic	Monthly
Climatic	Average temperature of the month	Dynamic	Monthly
	Climate factor	Dynamic	Monthly

Table 4.2-1. Selected characteristic variables.

Table 4.2-1 consists of four fields whose headings are: ***type of variable***, ***name of the variable***, ***property*** and ***measurement***. The ***type of variable*** field allows dividing the set of variables into groups according to the function they fulfill within the methodology. The field ***name of the variable*** allows to identify each variable used. The ***property*** field allows the identification of whether the variable is static or has a changing behavior over time. Finally, the ***measurement*** field has the value "fixed" if the variable is static, since it does not change and for the dynamic variables indicates the periodicity of the measurement of them.

4.2.1.1 Detailed explanation of the variables

Next, a description of the selected variables will be made, explaining briefly the reasons that justify their inclusion within the proposed methodology.

- **User Identification Number (NIS):** This is the only variable whose inclusion is mandatory, since it allows to differentiate each user of the system and to map each variable of them.

- **User tariff:** It allows to establish to which type each user belongs, which confers the ability to eliminate all users that do not belong to the study set, which is composed of residential, commercial and industrial users. In addition, it allows divides between users of each type. For example, the existence of tariffs one to six (1-6) in residential users. This is important because it is known that in slums (low stratum) there is a marked tendency for fraud to occur.
- **User measurement equipment:** It is used to identify the type and brand of the meter of each user. This is because it is known that the more modern and sophisticated the meter, the more difficult it is to manipulate it to commit fraud. On the contrary, those older meters are constantly violated for this purpose.
- **Geographic location of the user:** The commercialization companies of the country divide the geographical space into localities, which facilitates the sectorization of the clients. This variable complements the information obtained from the tariff, since it allows to analyze in detail the behavior of smaller groups of users than those obtained with that variable. The inclusion of this variable is justified because it is known that there are localities where the level of occurrence of fraud is considerably higher than others.
- **Consumption profile:** It is considered the most important variable within the methodology, since it allows to know the historical record of consumption of each client during the study window of the system. It is remarkable that it is the variable that more information contributes on the consumption behavior of the users.
- **Meter reading:** It is used to set the consumption and billing cycles of each customer. The inclusion of this variable allows the calculation of the days billed per period of each user, information that allows explaining changes in consumption behaviors that are not related to fraud.
- **Number of complaints made:** The inclusion of this variable is justified because there is a high correlation between the complaints made and the

client's behavior. It is clear that a client who complaints has a high probability of having a normal behavior, this because to complain is exposed to receive an inspection visit at home; which would not be convenient for a customer who is committing fraud.

- **Number of overdue bills:** Similar to the previous case, there is a high correlation between the amount of debt and the behavior of the user. The greater the number of overdue bills, the greater the probability that the user will commit fraud to evade or decrease the payment of the same.
- **Customer fraud history:** It is used to know if during its history of consumption, a user has been discovered committing fraud. This is due to the fact that there is a marked tendency to repeat for this type of user.
- **Reading and billing anomalies:** It explains changes in consumption behavior that are not associated with fraudulent conduct. This gives the system a high degree of robustness, since it gives the ability to recognize when a change in consumption behavior is due to an error in reading or billing and not because of the incidence of fraudulent behavior.
- **Balance and moorage of macrometers:** These variables conform the proposal of this research to determine when a user is normal (non-fraudulent). There is currently no direct way beyond an inspection to establish whether a customer has a normal consumption behavior, which was a problem for the execution of the system since the classification stage (stage 3) is supervised type. The above involved the need to have a set of users whose normal behaviors were checked, to use them in the training of the algorithm.

  Therefore, the advantage of the macromeasuring method used by commercialization companies (see Chapter 1, Section 1.1) was used to establish a criterion for estimating whether a user has normal behavior. Balance is the consumption presented by a macrometer, while moorage is the individual consumption of each of the meters (child meters) connected to that macrometer. The established criterion is to add the consumption of each child meter and compare them with the consumption of the associated macrometer, if the difference does not exceed 5% it is possible to ensure

that users of said macrometer can be considered as normal. Otherwise, the criterion allows to know the existence of fraud. However, it is not possible to know which of the users are fraudulent and which are not. For this reason, these variables will only be used to obtain the set of normal users for training.

- **Results of campaigns and inspections:** Unlike the previous case, obtaining the set of fraudulent users for training the system is quite simple. This is because this variable is the result of the inspections done to the clients, which indicate if the customer was found to be committing fraud or not. In case of being found as fraudulent can already be used as example data to train the system.
- **Average temperature of the month:** This variable allows explaining changes in consumer consumption behaviors. This is due to the fact that temperature is known to be one of the determining factors of people's energy consumption habits. It is clear that in times of high temperatures, users will increase consumption due to the continuous operation of air conditioning equipment (fans and air conditioners).
- **Climate factor:** It is a variable proposed in this research that allows to explain the existence of phenomena of “El Niño” and “La Niña”. These are phenomena that modify massively the normal climatic conditions of the region, bringing with it notorious changes of the consumption behaviors of the users.

4.2.2. Group 2: Conditioning and execution of the intelligent system

This group contains methodological steps from two to four (2-4), through which the conditioning and the execution of the three stages of the intelligent system are carried out. The first stage is an unsupervised grouping of similar consumption profiles, the second is the modeling and prediction of future consumption, and the third is a fraudulent user detector that receives as input the outputs of the previous stages, together with the other variables already described in Section 4.2.1.

As mentioned above, one of the innovations presented by this proposal is that the first two stages of the system are independent of each other, which allows them to be executed separately and subsequently the outputs of these are used for the execution of the third stage. As a result of this approach is obtained the structure of steps two to four of the methodology developed in this research (see figure 4.2-2). In which it is observed that step two corresponds to the execution of the first two stages of the system, where the branching allows explaining that these stages are independent of each other, being able to be conditioned, executed and validated individually.

Once the results of the first two stages are obtained, step 3 of the methodology is continued, which corresponds to the conditioning of all the variables that will enter to the final stage (stage 3). Finally, arriving at step four, the variables already conditioned are entered to the detector and this one is executed to carry out the classification of the users in fraudulent and non-fraudulent.

Having explained each of the steps that compose the second group of the proposed methodology, the block diagram of the intelligent system is presented (see Figure 4.2-3). The most significant and important contribution of this research is concentrated in the second group of the methodology (steps 2 to 4), since to carry out these steps first the design and implementation of the intelligent system must be carried out. This corresponds to the objectives proposed at the beginning of the research, placing this project within the Grupo de Investigación en Robótica y Sistemas Inteligentes (GIRSI) and its line of research in intelligent systems.

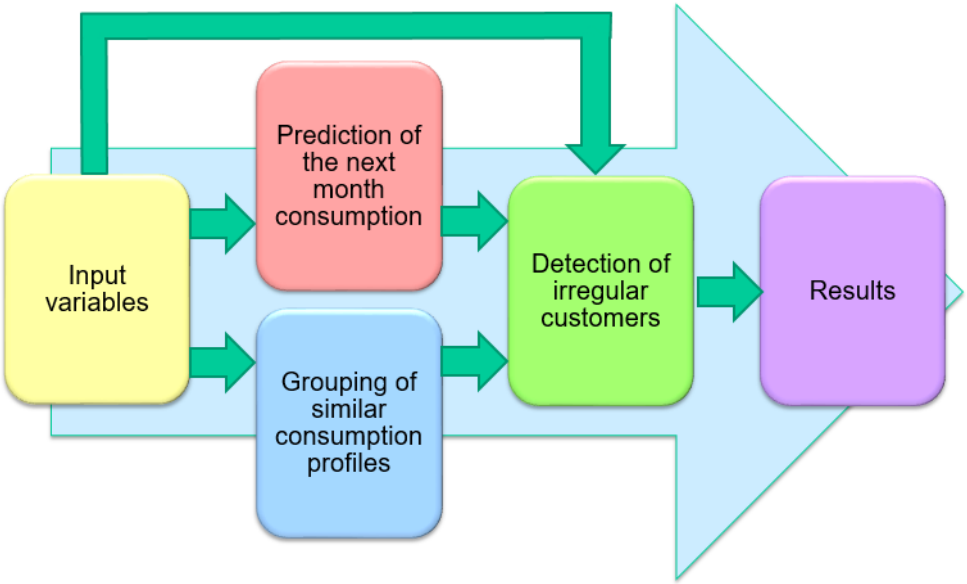


Fig 4.2-3. Proposed intelligent system block diagram.

From figure 4.2-3 it can be seen that the presented block diagram corresponds to the description previously made, the grouping and prediction stages are independent and their outputs are entered in the detection stage. Additionally, it can be seen that a portion of the inputs is used to execute said stages (1 and 2) and the remaining variables are entered as inputs of stage three (detection). The detailed description of each of the stages of the system and their respective functions are presented below.

4.2.2.1. Intelligent detection of fraudulent users using intelligent reinforcement stages

From the review carried out in Chapter 3, it was possible to conclude that in those works where more than one computational intelligence technique is used, the intention of the authors in the majority of the cases was centered in the implementation of hybrid algorithms to obtain better performances that using only one intelligent algorithm. In these cases, the result is a one-stage system but with superior functionality. In other cases, where the use of more than one technique was focused on obtaining a system with multiple stages, these were executed in cascade to sequentially improve the performance of the system.

In the case of this proposal, it was decided to rethink the situation and focus the design of the system from a different perspective. In which the use of various computational intelligence techniques allowed to obtain a system with multiple stages but with a different operating principle. The operation would no longer be to improve the performance by means of a sequential execution of several stages, but to execute several independent stages whose outputs served as reinforcement inputs in the detector, those outputs would accompany the initial set of inputs in the same. The purpose of this approach is to improve performance by obtaining additional inputs from those obtained from commercialization companies. The additional inputs would allow to reinforce the detection because these are the product of computational intelligence techniques, focused on characterizing aspects not measurable but highly significant in the analysis of electrical energy consumption behaviors. This is due to the ability of these algorithms to generalize and recognize patterns, which would lead to the obtaining of variables that would have a higher weight because they are the result of an intelligent analysis, allowing a better and more detailed characterization of the behavior of the users.

As a result, the addition of two reinforcement inputs to the detection is proposed, each one is a product of the execution of its corresponding stage based on computational intelligence techniques. The first input is the result of the grouping stage, while the second input is the product of the prediction stage. Both stages are described below.

4.2.2.2. Stage 1: Unsupervised grouping of consumption profiles

The reason for the development of this stage lies in the following assumption, ***if it is possible to divide the set of users into groups according to their consumption behaviors, so that in each group result users with similar behaviors among themselves. Then the users of the same group will share the same characteristics and patterns, whether these indicators of fraudulent behavior or not***. Therefore, the grouping is aimed at capturing the aspects shared by the members of each group, based on these to generalize about the behavior of a user from the group to which it belongs.

This stage corresponds to the unsupervised learning approach, because the group to which each user belongs is not previously known. It is the duty of the computational intelligence technique to find the possible groups that can be obtained from an intrinsic analysis of the input variables. Which should result in that the users of the same group must have a similar consumption behavior, while differentiating as much as possible of the users of the other groups.

As mentioned in Chapter 2, the process described above is carried out by the use of clustering algorithms. From the set of input variables presented in Section 4.2.1, the only one that will be used for the execution of this stage is the consumption profile of the users, which is the time series that represents the consumption behavior of them within the study window.

It is important to add that since the function of this stage is to group according to similar consumption behaviors, it is necessary to ensure that the input variable allows the algorithm to fulfill such function. This is why it was decided to normalize each of the consumption curves (profiles) of the users with respect to their individual maximum value, so as to eliminate the affectation that introduces the magnitude of consumption and it is possible to capture only the behavior of the same. This is exemplified by the two consumption profiles shown in Figure 4.2-4, where it is possible to observe that from the image (a) one could conclude that both profiles have different consumption behaviors. However, after normalizing, the curves of the image (b) are obtained, where it is possible to see that they have similar consumption behaviors.

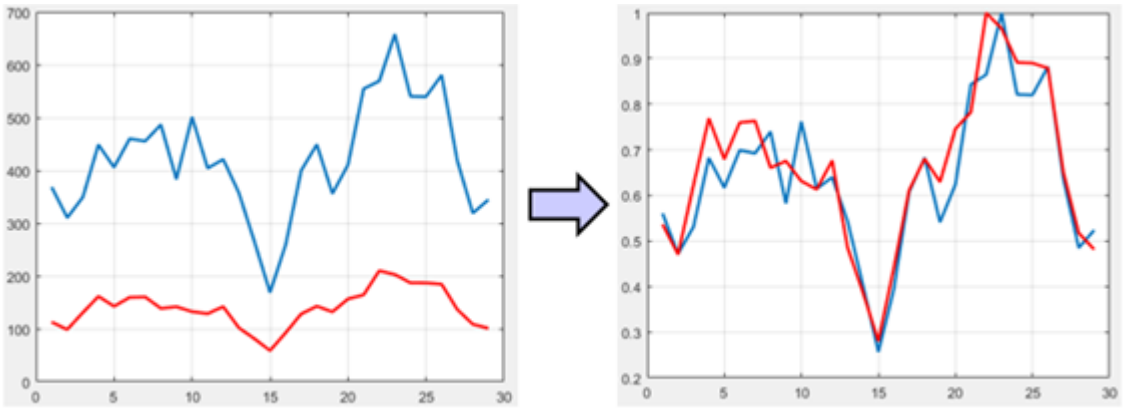


Fig 4.2-4. Example of the normalization of consumption profiles of step 1.

An example of the implemented clustering stage is presented in Figure 4.2-5, where the user profiles are entered into the algorithm and this is responsible for the grouping according to the degree of similarity present in them. The result is a set of groups (6) where the profiles located in each one of these groups present consumption behaviors (shape of the curve) similar to each other and different from those profiles located in other groups.

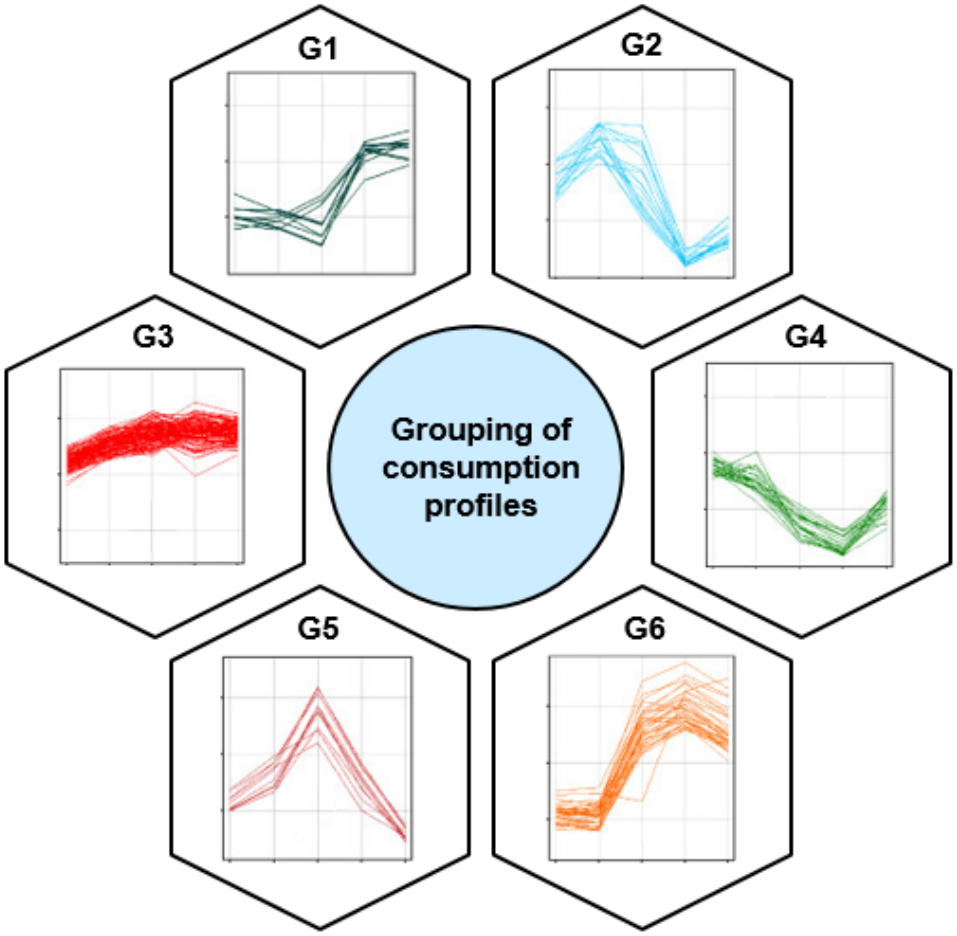


Fig 4.2-5. Illustrated example of the operation of stage 1.

The output of the consumption profile grouping stage is a number for each user of the system, that number indicates in which group each customer was located after the execution of the stage. This output is one of the additional reinforcement inputs that will be entered to the detection (step 3). The complete process and its performance metrics are explained in more detail in the next chapter.

4.2.2.3. Stage 2: Consumption profiles modeling and prediction of the next month consumption

Similar to the previous stage, the approach of this stage is also based on an assumption, which is stated below. ***If it is possible to find a model that faithfully represents a user's consumption profile, this model can be used to predict the behavior of this user in the future***. More specifically, this stage seeks to find the best model to represent the consumption profile of each user of the system. So that it is possible to perform with a high degree of certainty, the prediction of the consumption of each user for the next month.

The idea behind the implementation of this stage is to be able to compare the prediction made with the actual value consumed by the user in that month. If the model used to represent and predict the user behavior is the best possible, the prediction will have the highest degree of certainty possible. This concept can be used to make the above mentioned comparison and expect the deviation between both values (predicted and real) to be minimal. If not, using the model that best represents the user, there would be reasons to conclude that this user did not behave as he should. Therefore, there has been a change in the user normal consumption behavior and there is a possibility that this may have engaged in fraudulent conduct.

From Chapter 2, it is possible to see that this stage corresponds to a regression problem, where a set of variables must be used to find the function that best allows to fit the known data of the output. In this context, some of the input variables of Section 4.2.1 must be used to find the model that best fit the consumption profile data of each user.

The variables taken into account as inputs to this stage are:

- Consumption profile of each user.
- Average temperature of the month.
- Climate factor of the month.
- User Tariff.
- Meter readings.

The variable ***Meter readings*** was not used directly, but was used to calculate the days billed to users in each period. Additionally, two additional inputs were used corresponding to the year and month of the period to be predicted, which due to being dates are not considered within the set of main variables of Section 4.2.1.

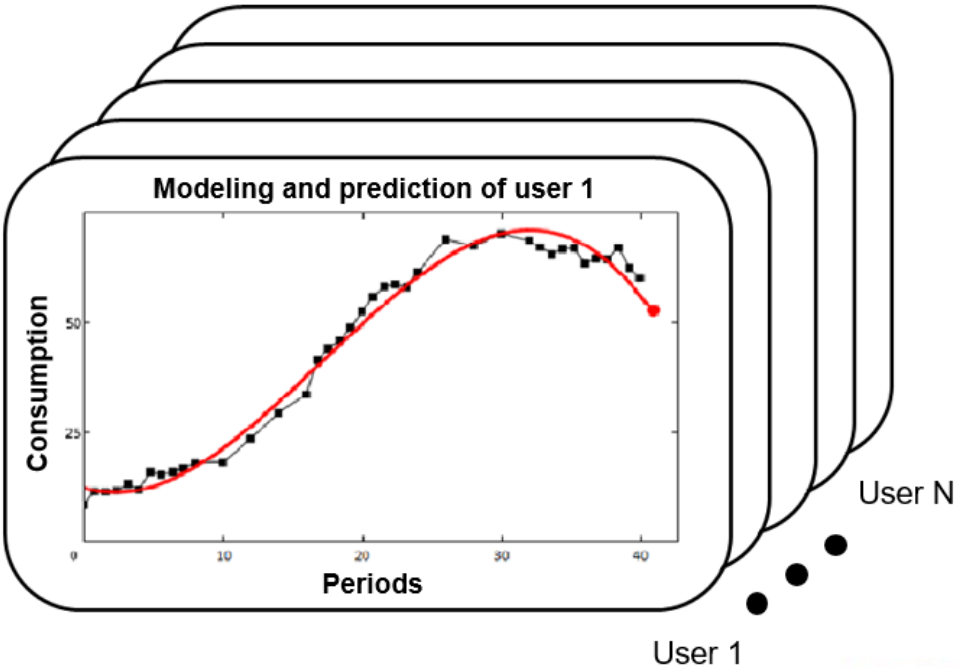


Fig 4.2-6. Illustrated example of the operation of stage 2.

Figure 4.2-6 represents the operation of the modeling and prediction stage. It is possible to observe that given a consumption profile of a user of the system (**black line with squares**) and the other variables mentioned for that user, the stage fits the model that best explains the consumption profile of this user (**red line**). Once this model is found, it is used to predict the consumption of the next month of the user under study (**red circle**). The difference between the value of the **red circle** (prediction) and the actual value (when it occurs) is the output of stage 2 for this user, which will be entered as additional reinforcement input in the detection (stage 3). This process is repeated independently for each of the ***N*** system users.

4.2.2.4. Stage 3: Detection of fraudulent users

This is the final stage of the proposed intelligent system, since it is at this point that the intelligent detection of fraudulent users is performed. It should be taken

into account that the better the performance of this stage, the better the system performance and the more satisfactorily the proposed objectives will be met.

From Chapter 2 it is possible to see that the design of this stage corresponds to the solution of a classification problem, where a set of variables must be used to obtain the function that allows to better classify the elements of the input set in their respective groups. In this context, we must use those variables of Section 4.2.1 that allow the algorithm to perform a suitable generalization of the characteristics that differentiate fraudulent users from non-fraudulent ones. In addition to these variables, the outputs of stages 1 and 2 of the system will also be entered as inputs to this stage, since these represent the reinforcement inputs whose purpose is to improve the characterization of the consumption behavior provided by the initially mentioned variables.

The variables taken into account as inputs to this stage are:

- User measurement equipment.
- Geographic location of the user.
- Number of complaints made
- Number of overdue bills.
- Customer fraud history.
- Reading and billing anomalies.
- Group to which each user belongs (output stage 1).
- Consumption deviation (output stage 2).

The reading and billing anomalies constitute a list of more than fifty items, however, for the use within the system these were divided into three large groups. The first group is made up of anomalies directed to unoccupied properties, which allow to explain changes in consumption that are the product of recently vacated properties. The second group consists of the anomalies that are not directed to frauds, which allow to explain changes in the consumption that are due to errors in the reading or billing. The third group corresponds to anomalies directed to fraud, which allow to explain changes in the consumption of users who have been recently discovered committing fraud. In this way the variable of reading and

billing anomalies was used, allowing to adapt it to the characterization of normal or fraudulent consumption behaviors.

Due to all input variables are of numeric type, for the entry of these to the classifier algorithm most of these were expressed in binary representation. Such transformation was carried out on those integer numerical variables, which in this case are all except for the ***deviation of consumption***, which can take values with decimal digits. This is due to the fact that the experience in the use of intelligent algorithms has shown that the inputs represented in binary form facilitate the recognition of patterns, allowing to obtain better results in the desired classification.

The classification problems belong to the supervised learning approach, therefore, the training of the intelligent technique is performed by examples with known results. To achieve the training of the intelligent classifier for the detection of fraudulent users, not only were the mentioned inputs required, but a set of known outputs were required for the combinations of said inputs. This is the function of the variables ***balance and moorage of macrometers***, as well as of the variable ***results of campaigns and inspections***. These allow to determine if a user has normal or fraudulent behavior, so that can be given to training the classifier with the examples determined by the sets of inputs and their outputs.

Figure 4.2-7 shows an example of the operation of the detection stage. In which it can be seen that initially a set of examples is presented, which correspond to a group of users whose input variables are known, as well as their corresponding output (normal or fraudulent). The data of these example users is entered for training the classification algorithm, which at the end of the process is able to determine if a user is committing fraud or not depending on the values of their inputs. The result is the stage of detection of fraudulent users, which at this point can already be used to classify a new set of users that are not part of the set of training examples. Of which its exit is not known and it is wanted to use said stage to determine it.

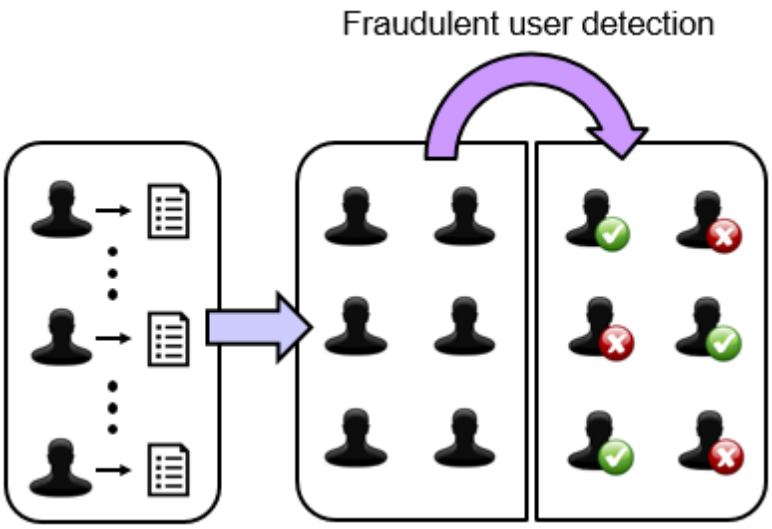


Fig 4.2-7. Illustrated example of the training and operation of stage 3.

4.2.3. Group 3: Results

This group contains the methodological step five (5), through which results analysis and validation of the detection stage is carried out. It corresponds to the evaluation of the intelligent system performance for detection of fraudulent users proposed in this research. At this point, the necessary tests are performed to verify the operation of the system as a tool to improve the process of reducing non-technical losses resulting from fraudulent connections. In this step, corrections and modifications that are considered necessary for the system improvement can be implemented.

Once satisfactory results are achieved and the established performance criteria are met, the system is validated and can be used to detect fraudulent users. This leads to the conclusion of the proposed methodology as a result of this research.

Chapter 5

Implementation of the intelligent system for detection of fraudulent users

This chapter presents the application of the proposed approach to the problem of non-technical losses of electrical energy that are the product of fraudulent connections, applied to the detection of the users that generate these losses. Some considerations are presented about the system and the set of clients used to test it. The characteristics of the variables and the implementation of the algorithms that compose the grouping of consumption profiles, profile modeling and prediction of consumption of the following month, as well as the detection of fraudulent users are described in detail.

In the previous chapter the product of this investigation was proposed, which is a methodology that allows to improve the process of detection of users with fraudulent connections. It was emphasized that the main and most important step in this methodology corresponds to the design and implementation of the intelligent system that allows this purpose to be met. The process of selecting the variables to be used in the system was described in detail, justifying the reasons that support the choice of each one of them. Finally, the conceptual design of the system and of each of its stages was presented, explaining the functionality and the contribution of these within it.

This chapter describes in detail the process of implementing the system and each of its stages, allowing the reader to get a complete view of the internal structure and operation of the same.

5.1. Generalities of the system

The concept of ***generalities*** allows to group all those characteristics that impose restrictions, set limitations and mark the scope in the operation of the proposed system. The following are those taken into account for the design, implementation and execution of the same.

5.1.1. Geographical space

The proposed approach is a general methodology that is composed of a series of steps whose correct execution allows to detect fraudulent users. Although the approach can be applied generally to any system of electric power distribution, a set of users of a specific place where to carry out the tests that allow to validate the system is needed. For convenience[1], the place where this research was developed was chosen as the geographical space for testing, this corresponds to the city of Barranquilla, capital of the Atlantico department.

After consulting with the electrical energy commercialization company of the region, it was possible to know that it divides the users of each department into groups known as ***delegations***. These delegations allow a more detailed study of the users within each department of the Colombian Caribbean region. In the case of Barranquilla, the site selected for testing the system, this is contained in the delegation known as ***Atlantico Norte***, which is also integrated by the municipalities of Malambo, Puerto Colombia and Soledad.

From the above, it was decided to extend the geographic space initially selected, so that it coincides with the delegation that manages the commercialization company. Therefore, it was finally established that the set of users on which the system would be tested would be those belonging to the municipalities of Barranquilla, Malambo, Puerto Colombia and Soledad. Figure 5.1-1 shows the political map of the Atlantico department, on which has been marked with a ***red*** line those municipalities whose customers will be part of the test base of the

[1] The term ***convenience*** is used to refer to the facility to obtain the information of the customer database, from which the variables necessary for the characterization of users in the system are extracted. This is possible due to the direct communication in an on-site manner with employees of the region's energy commercialization company.

intelligent system for detecting fraudulent users.

Fig 5.1-1. Map of the Atlantico department with the demarcation of the Atlantico Norte delegation. (www.atlantico.gov.co)

5.1.2. Time window

It has been mentioned repetitively throughout the document that the performance of intelligent algorithms is highly dependent on the ***quantity*** and ***quality*** of the information that is supplied to them. The time window is the feature that directly defines the ***quantity*** of information on which the proposed system will run. It should be noted that the study of consumption behavior of electrical energy users is the analysis of a process that develops in the time domain. Therefore, fixing the time window of study directly impacts the results that will be obtained from the analysis; to a larger study window, more information can be obtained and the analysis will be more complete.

At the beginning of the thesis it was possible to fix the end of the window of study time, so that it coincided with the beginning of the investigation. This allows defining the end of the study window in the month of July of the year 2015. However, setting the window start was not easy because there was no specific time to take as a reference, and the ***quantity*** of the information should be guaranteed by means of a correct selection of the window extension. This problem was solved by contacting the commercialization company of the region, where it was possible to know that the information available in the database for all the selected characterization variables (see chapter 4, section 4.2.1) exists from the year 2010 onwards.

After receiving the information of the customer database provided by the commercialization company (year 2010 onwards), conducting an inspection of the data it was possible to establish that the ***quality*** of the information could be considered acceptable since 2011. The above due that in the data of the year 2010 there was a large amount of incomplete and erroneous data, which would introduce a considerable level of error when executing the intelligent algorithms. As a result, the start of the study time window was fixed in the month of January 2011, which indicates that there was a window of fifty-five (55) periods (months) for the execution of the system.

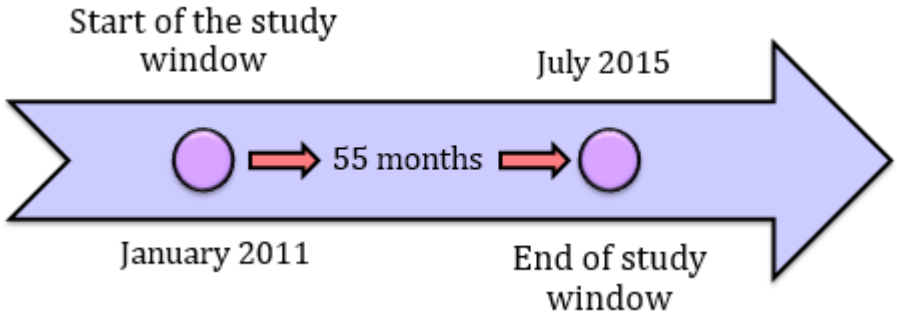


Fig 5.1-2. Timeline with system study window.

5.1.3. Types of users

In the previous chapter, when the variable ***User tariff*** was described (see section 4.2.1.1), it was mentioned that the users on whom the system would be run had to belong to one of the valid user types, which are residential, commercial and

industrial. This is due to the great majority of users of the electricity sector belong to one of these three major types. There are other types used by commercialization companies, however, as they represent a small minority have not been taken into account within this proposal.

Within the valid groups of users are presented divisions that allow companies to bill the electric power service differently to each user. For purposes of this approach the divisions mentioned will only be taken into account for residential users, this because the consumption behavior of these consumers can be better characterized by such divisions. In the case of commercial and industrial it is assumed that consumption behavior among users of the same type is not affected by this factor. The above is summarized in the following table:

Initial type	Assumed type
Residential stratum 1	Residential stratum 1
Residential stratum 2	Residential stratum 2
Residential stratum 3	Residential stratum 3
Residential stratum 4	Residential stratum 4
Residential stratum 5	Residential stratum 5
Residential stratum 6	Residential stratum 6
Commercial type 1	Commercial
Commercial type 2	
Commercial type 3	
Industrial type 1	Industrial
Industrial type 2	
Other types	Eliminated

Table 5.1-1. Types of initial users and types of assumed users.

5.1.4. Users under macrometering

This characteristic was mentioned when the proposal of this investigation to characterize the users with normal behaviors of consumption (non-fraudulent) was described. It was mentioned that a criterion was put in place that allowed to establish when the users connected to a macrometer could be considered normal (see chapter 4, section 4.2.1.1). This was necessary because of the supervised nature of the classification problem that represents the detection of fraudulent users. Where initially the system should be trained through a set of examples

whose answers are known, purpose for which the proposed criterion derived from macrometering was used.

Taking advantage of the macrometering technique used by commercialization companies imposes a limitation on the system. This is reflected in the obligation to ensure that all users of the system are under the action of a macrometer, this in order to be able to perform the validation to each and every one to see if they can be cataloged as normal. Therefore, the system will only be executed on users that meet this condition, those that do not comply will not be taken into account.

5.1.5. Pre-filtering processes

Once the general characteristics of the system have been described, it is necessary to mention that they serve as a first step in conditioning the information. This is because they allow the filtering of the users on which the system will finally run. The customer database received from the commercialization company consists of the total users that it serves in the Caribbean region of Colombia, in order to obtain the set of users that will be part of the system, a series of filters are made according to the general characteristics described. These are:

- Elimination of all users that do not belong to the Atlantico Norte delegation.
- Elimination of users that do not belong to any of the three valid types.
- Elimination of users that are not connected to a macrometer.
- Elimination of users presenting incomplete information for any of the characterization variables (see chapter 4, section 4.2.1) during the study window.

In addition to the aforementioned filters, others were also added to guarantee the quality of the information contained in the most important variable of the system, which is the ***consumption profile*** of the customers. Filters based on this variable are listed below:

- Elimination of users with any negative consumption.
- Elimination of users with any zero consumption.

The diagram shown in Figure 5.1-3 illustrates the pre-filtering process that is performed on the database with the customer variables received from the region's electricity commercialization company. It shows the total number of clients and the amount of them that is resulting after applying each of the described filters. It is observed that the final number of users on which the system will be executed is 92,794.

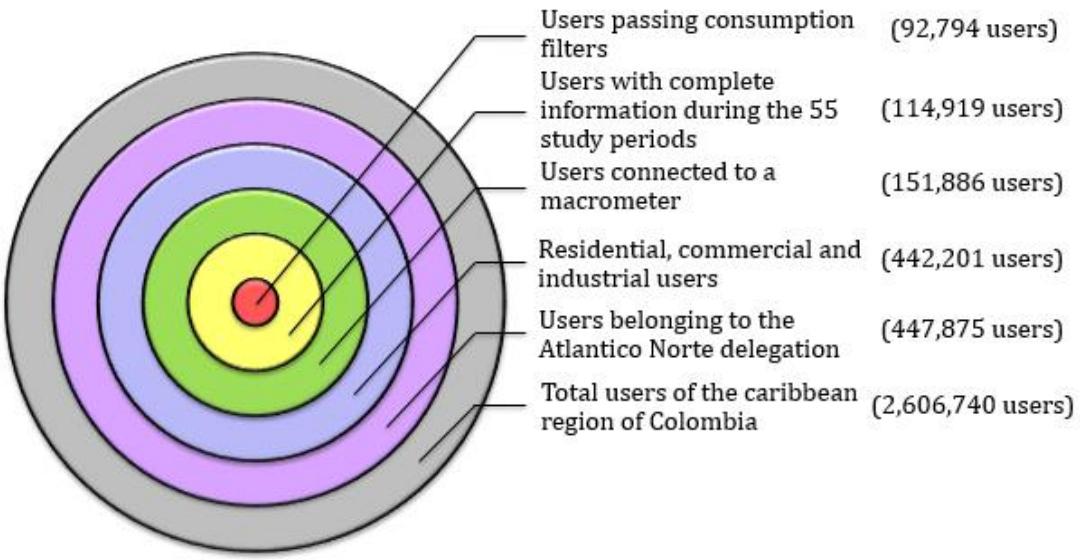


Fig 5.1-3. Application of the filters and the resulting number of users for each of these.

5.1.6. Software used for system implementation

The development environment selected was Matlab, because it is the most complete numerical computing platform, which greatly facilitates the development of the codes that compose the algorithms of the system. Additionally, it should be noted that of the environments with the numerical computing capabilities required by this system, the university only has the Matlab license. The reasons given make this software the ideal choice for the development of the computational part of this research.

It is necessary to clarify that the codes of the techniques of statistical modeling and computational intelligence techniques used were not developed during the course of the research. These are already included as functions of Matlab that can be invoked by the user when they are needed.

5.1.7. Periodicity of system execution

The intelligent system for detecting fraudulent users is designed to run monthly (once a month), allowing to detect customers with frauds every month. From the previous approach arises the need to have the information of the input variables for all the periods (months) that compose the study window.

5.2. Implementation of the stage of unsupervised grouping of consumption profiles (stage 1)

As explained in chapter 4, the goal behind the implementation of this step is to obtain one of the reinforcement variables for the detector (stage 3). The function of the cluster is to divide the set of users into groups, so that the users of the same group exhibit similar consumption behaviors, and differ from the behavior of the users present in the other groups. The output of this stage is the reinforcement variable that captures the patterns and characteristics that are common among the members of each group, allowing to generalize about the behavior of a user from the knowledge of the group where it is located.

5.2.1. Input variables

This step uses only a single input variable of all those described in section 4.2.1, which is the consumption profile of each of the users of the system. It should be remembered that because the study window is of fifty-five months, the consumption profiles of each user will be time series with equal number of values.

5.2.2. Conditioning of input variables

The conditioning of the input variable for this stage corresponds to a transformation, in which the normalization of the consumption profiles of each user is carried out with respect to the greater consumption of each one. That is, for each user of the system, its highest energy consumption value is located within the study window, and then dividing all the consumption values of its profile into that maximum value. This leads to a series with values in the range of zero to one, allowing to capture only the consumption behavior of the profile (shape of the

curve) by eliminating the scaling factor provided by the magnitude (see chapter 4, section 4.2.2.2).

5.2.3. Structure of the stage

5.2.3.1 Clustering algorithm

In the description of the functionality of this stage, which took place in section 4.2.2.2, it was clarified that to carry out the unsupervised grouping a clustering technique should be used. The two most widely used techniques for solving clustering problems are the self-organizing maps (SOM) and K-Means. For the implementation of this stage it was decided to use SOM because to configure it is necessary to adjust four parameters, unlike K-Means whose configuration only depends on two parameters. This is based on the ability to explore a larger space of solutions that provides the use of a greater number of parameters. Exploring as many solutions as possible allows a better response and guarantees that there is no better solution than the one obtained, which validates the selection made.

The parameters that allow to configure a SOM are:

- **Dimension:** It corresponds to the total number of neurons on the map. It is directly related to the total number of groups that wish to be obtained from the grouping.
- **Topology:** It refers to the shape that results with the interconnection of the neurons on the map. This can be square, hexagonal or random.
- **Distance function:** It is the expression used to calculate the membership of each element to any of the groups in the network.
- **Number of training steps:** It corresponds to the number of iterations made during the mapping that this technique makes of the space covered by the elements to be grouped.

5.2.3.2 Optimizing the clustering using a search algorithm

Each of the described parameters can take different values, which allows to infer that the resulting number of combinations is a function of said quantity of values. In order for this step to improve the results of the detection of fraudulent users, it

is necessary to perform the best possible grouping, for which it is necessary to know what combination of parameters allows such clustering. This indicates that the entire solution space of possible combinations must be explored, this is a great effort and would lead to the realization of an experimental design with a high number of tests. In order to overcome this obstacle, it was decided to use a search algorithm, since this type of techniques allows finding the best possible solution (or getting too close to it) without the need of an experimental design.

There are a lot of search algorithms, however, currently stand out a set of techniques of optimization and search known as metaheuristics. These are mostly algorithms inspired by nature, which have proven to be highly efficient in solving these types of problems. The metaheuristic technique selected was the genetic algorithm, this being the only one that is natively included in Matlab.

The genetic algorithm represents each combination of the SOM parameters as a chromosome of its search space, this idea is represented by the illustration in figure 5.2-1. Ultimately, each chromosome of the genetic algorithm is just a vector containing a specific configuration of the SOM parameters.

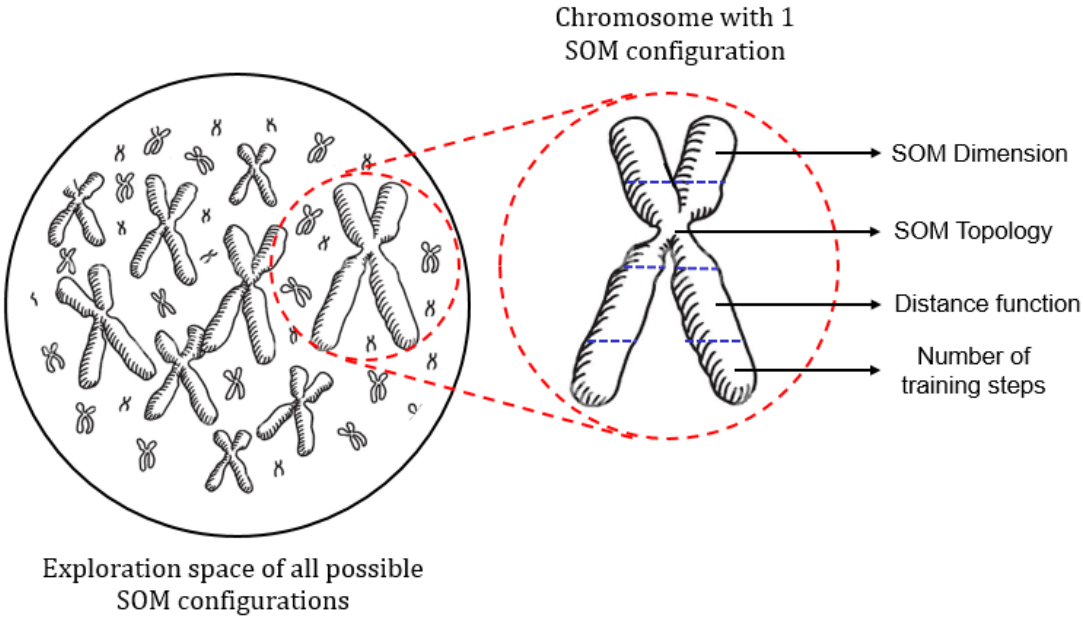


Fig 5.2-1. Representation of the configuration vector with the SOM parameters from the GA approach. (modified from www.istockphoto.com)

5.2.3.3 Functional structure of the stage

The previous approach allows defining the functional structure of the stage of unsupervised grouping of consumption profiles. Which turns out to be self-organization map within a genetic algorithm. The SOM is responsible for grouping consumption profiles for a given configuration of its parameters, while the genetic algorithm explores the combinations of possible parameter configurations until it finds the one that yields the best clustering. During the operation of this stage the GA will follow its operating principle to select the combinations of parameters with which the SOM will execute, this process will take place until it stops in the best solution found, that is to say, the best grouping. The functional structure of the stage is presented in Figure 5.2-2.

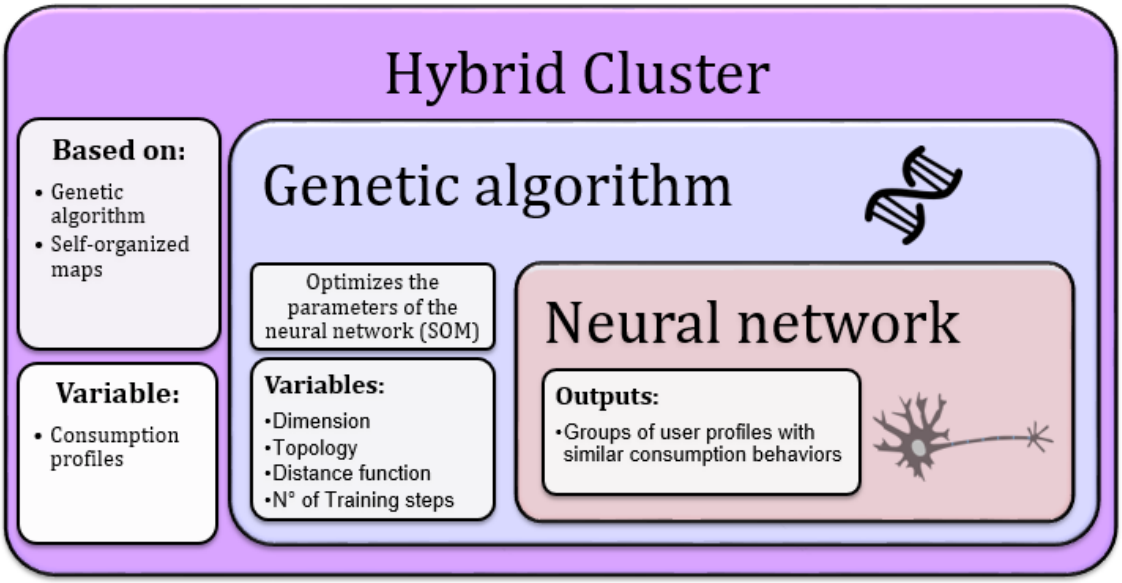


Fig 5.2-2. Functional structure of stage 1 of the system.

5.2.4. Conditions of execution

The expression ***conditions of execution*** refers to the division of the input dataset to obtain the training and test sets. For the case of this stage, because it is desired to obtain the best grouping of the total set of users of the system; all of them will be used for both the training and testing phases. This will remain the same for each and every one of the cluster executions until achieving the best grouping. It should be remembered that the intelligent algorithm (SOM) will be executed successively

while the GA finds the configuration of the one that yields the best results, and in each execution the training and testing phases take place. In conclusion, in each execution the SOM is trained and tested with the consumption profiles of the 92,794 users.

5.2.5. Implementation in Matlab

For the implementation in the selected development environment the values that each SOM parameter can take are defined, this in order to establish the total set of combinations that the genetic algorithm should explore. Additionally, the genetic algorithm is configured for the correct solution of the optimization problem that represents the best clustering of the SOM.

5.2.5.1 Self-organization map parameters

Below, the possible values of the SOM parameters are defined.

- **Dimension:** The minimum number of groups to obtain is four (4) and the maximum number is two hundred and fifty-six (256). It should be remembered that the SOM is always a square two-dimensional network (see chapter 2, section 2.3.3.3), so to obtain four groups a network of 2 x 2 neurons is generated and to obtain two hundred and fifty-six a network of 16 x 16 neurons is generated. The smallest number of groups that Matlab allows is four, so that will be the value of the lower bound. The selection of the upper limit value is explained in section 5.2.5.2.
- **Topology:** The three options that Matlab offers for this parameter are square, hexagonal or random topologies; all will be tried.
- **Distance function:** The four options that Matlab offers for this parameter will be tested, which are ‘dist’, ‘linkdist, ‘mandist’ y ‘boxdist’.
- **Number of training steps:** Three values will be tested, which are 50 steps, 100 steps and 150 steps.

From the above it is possible to see that there are fifteen values for the first parameter (2 to 16 for being two-dimensional), three for the second, four for the third and three for the fourth. This allows to know that there are a total of three

hundred sixty (360) possible combinations of SOM parameters. It is up to the genetic algorithm to find the one that yields the best grouping without having to evaluate them all, this is the idea behind the metaheuristic techniques.

For the use of Matlab GA function it became necessary to encode the values of the parameters in numerical representation. The values of the parameters ***Dimension*** and ***Number of training steps*** are already numerically expressed, therefore, only the parameters ***Topology*** and ***Distance function*** must be transformed. The result is presented in the following table:

Parameter	Original value	Value in numerical representation
Topology	Hexagonal	1
	Grid	2
	Random	3
Distance function	'dist'	1
	'linkdist'	2
	'mandist'	3
	'boxdist'	4

Table 5.2-1. Numeric representation of the parameters with text values.

5.2.5.2 Genetic Algorithm Configuration

This section presents the aspects taken into account for the adequacy of GA.

- **Multi-objective optimization:** Problems of grouping often bring with them an obstacle, which is explained by the following statement: ***The perfect way to group elements with different characteristics is through individualization***. That is, when you have different elements, perfect grouping occurs when you create "groups" of a single element. This is supported by the fact that in that case, the performance of the grouping is 100%. This is because when there is one element per group, the similarity of that element to itself is perfect.

  The idea of the previous paragraph represents a problem for the execution of the grouping stage, due to the desire to achieve the best grouping without falling into individualization. The solution implemented to avoid such problem is to use a multi-objective optimization in the genetic algorithm,

where the first objective function is to minimize the number of groups and the second objective function is to maximize clustering performance. The value of the first objective function is obtained directly from the evaluation of the ***dimension*** parameter, while the second corresponds to the performance metric of the SOM and will be presented later.

At this point it is already possible to explain that the value of the upper limit (set at 16) for the ***Dimension*** parameter was obtained experimentally, in order to avoid individualization in the clustering. For this, a criterion was set by which a result is accepted as valid if for the group with the least elements, they exceed 5% of the group with the highest number of elements. Resulting 16 x 16 (256 groups) as the maximum value of the SOM dimension for which the proposed criterion is met. The tests performed are presented in the next chapter.

- **Integer optimization:** From section 5.2.5.1 it can be seen that in their numerical representation, the values of the parameters only take integer values. The above was a problem because the genetic algorithm of Matlab for multi-objective problems works on values with decimal numbers, which are generated during the processes of crossover and mutation of the same. As a solution, a modification was made to the Matlab function that allowed to execute it on problems with integer parameters.
- **Cache Implementation:** Another problem with the Matlab genetic algorithm is that it does not keep a record of the chromosomes that have already been tested, so it falls into the repetitive evaluation of already explored solutions. Due to the number of users that must handle this stage, the above represents a considerable increase in the execution times of the same. This was solved by implementing an external cache for the Matlab GA, which allowed it to keep track of the already tested chromosomes so as not to evaluate them again. The result was a significant decrease in the execution time of the stage.

5.2.5.3 Cluster performance metrics

There are two indexes accepted by the scientific community for the performance evaluation of clustering algorithms, the Davies - Bouldin index and the Dunn index (Mary & Jebarajan, 2014). However, due to the poor results obtained through its application to this particular problem, it was decided to propose a metric of performance according to the problem of grouping similar consumption profiles. The explanation of the proposed metric will be presented by an illustrative example of the ideal case, to later extend it to real cases.

The problem to be solved is to group profiles with similar consumption behaviors, this means that the shapes of consumption curves between members of the same group should be as similar as possible. Assume that stage 1 of the system has been used to perform a grouping of a set of users into ***N*** groups and that user consumption profiles consist of twenty-four periods (24). The results obtained at the exit of the stage are shown in Figure 5.2-3.

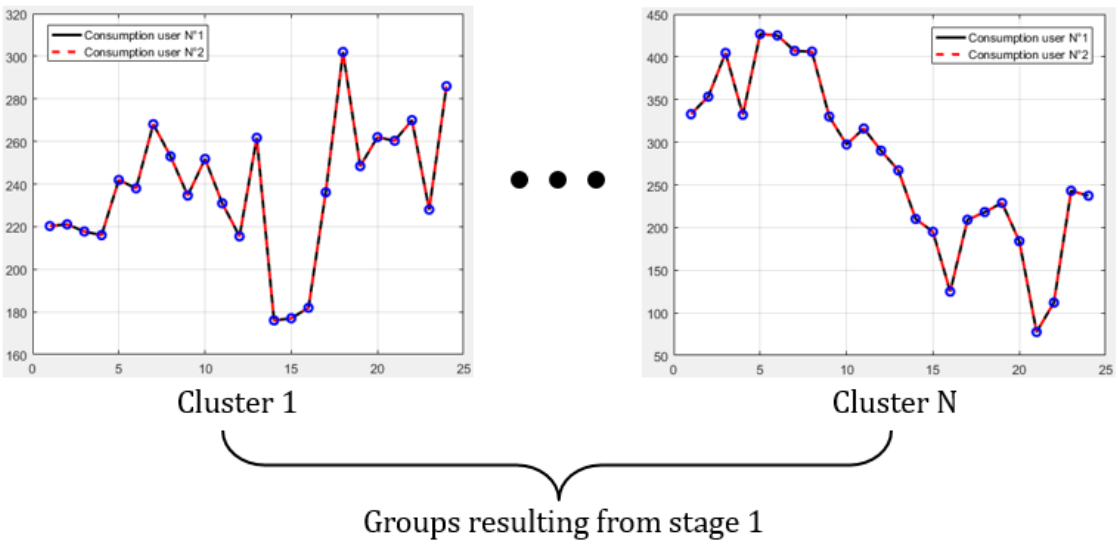


Fig 5.2-3. Example of the ideal case of the execution of step 1.

Observing the graph corresponding to the cluster (group) 1, it is possible to notice that two profiles were located in this one after the grouping (***black*** line and dotted ***red*** line). It is clear that the process was carried out correctly for that group since both profiles are ***perfectly*** similar (they are superimposed). If we calculate the standard deviation of the two profiles for period 1, we obtain zero as result, since

they presented equal consumption in that period. It is possible to realize that this is repeated in the same way for periods from 2 to 24 because the curves are exactly the same.

$$\sigma_1 = 0,\ \sigma_2 = 0, \cdots,\ \sigma_{24} = 0$$

If the average of the deviations obtained in each period is calculated, a zero result will be obtained, which indicates that the grouping was the best possible due to the perfect similarity of the two curves.

$$\bar{c}_1 = 0$$

Extending the approach for all the groups resulting from the process, we arrive at the group N, where it can be observed that the same case described for group 1 is presented. Therefore, the average of the deviations by period of the group N is also zero.

$$\bar{c}_N = 0$$

Assuming that the same situation occurred in all N groups, the average of the deviations per period for each group will be zero.

$$\bar{c}_1 = 0, \bar{c}_2 = 0, \cdots, \bar{c}_N = 0$$

Calculating the mean of the averages of each cluster will result in zero.

$$\bar{C}_T = 0$$

It can be concluded that this example corresponds to the best grouping possible, where the resulting curves in each group are perfectly similar to each other. Yielding as a result a total average $\bar{C}_T$ of zero. Now assume that the curves of the resulting profiles in each group are not perfectly similar. Extending the concept described it is possible to infer that the grouping is better when for a group the standard deviations in each period are minimal, since this will mean that the average of the deviations of that group is also minimum. If this occurs for all $\boldsymbol{N}$ groups, then the average of the deviations will be minimal for all groups, leading to the overall average of the process being minimal as well. The value of the total average $\bar{C}_T$ is the metric proposed to measure the performance of stage 1 of the

system. Becoming the value of the second objective function of the genetic algorithm.

5.3. Implementation of the stage of consumption profiles modeling and prediction of the next month consumption (stage 2)

According to the presented in chapter 4, the approach of this stage allows to obtain the second reinforcement variable that will be used to improve the classification of users that will take place in stage 3 (detector). The function of this part of the system is to find for each user the model that allows to better explain their consumption profile, so that it is possible to make a correct prediction of consumption of these in the following period. For each user, the output of this stage is the difference between the predicted value and the actual value. This is because a good model is able to faithfully represent the behavior of each user, which allows to expect that the difference between both values is small. In the opposite case, it can be ensured that a behavioral change has occurred and therefore, there is the possibility of the occurrence of fraud. From now it will refer to the difference between the actual and predicted values as ***consumption deviation.***

5.3.1. Input variables

Due to this step analyzes each user individually, the input variables that are required must be included for all the users of the system. These variables are mentioned below:

- User consumption profile (series of 55 periods).
- Average temperature of the month (for 55 periods).
- Climate factor of the month (for 55 periods).
- User tariff (fixed).
- Billed days of the month (for 55 periods).

It is necessary to emphasize that, in the majority the presented variables are obtained from the database supplied by the commercialization company of the region. The exception to this are the variables ***Average temperature of the month***

and ***Climate factor of the month***. These were obtained from the meteorological station of the Ernesto Cortissoz airport and the Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) respectively (wunderground, 2017; Instituto de Hidrología, Meteorología y Estudios Ambientales, 2016).

5.3.2. Conditioning of input variables

For the data corresponding to the input variables of this stage no conditioning or transformation was performed. This is due to the fact that this part of the system is executed independently for each one of the users of the same. Therefore, when evaluating each user individually, the data can be entered directly because they do not depend on the information of another client.

5.3.3. Structure of the stage

5.3.3.1 Part 1: Obtaining the statistical model

There are a great number of techniques that allow to obtain models for fitting of time series, standing out among all the others are those that are product of statistical analysis, as well as those that result from the use of intelligent algorithms. At the beginning of the investigation, it was proposed to implement this stage using only computational intelligence techniques, however, after multiple and successive tests, it was concluded that the consumption profiles were not easily characterized by intelligent algorithms.

An important fact is that, although the computational intelligence techniques are able to solve a wide variety of problems, it must be kept in mind that these belong to a paradigm known as soft-computing. In which the benefits of numerical computing are exploited, where there is nothing analytical and, on the contrary, one lives in the world of approximations. The above is mentioned to comment that in many occasions, the goodness of the intelligent algorithms make them the first option to solve any problem. However, it tends to be forgotten that for many problems there are analytical models whose results are already proven. It should be kept in mind that the potential of computational intelligence techniques lies in the ability to obtain process models that are not analytically characterizable.

Therefore, if an analytical model exists for the solution of a problem, it must be the option to be used unless it is computationally expensive.

As mentioned in the beginning of this section and in agreement with the mentioned in the previous paragraph, for the representation of time series exist statistical models whose operation is already proven. This allowed the decision to use these analytical models as the basis for the development of this stage. It was defined that the models to be used for the analysis of consumption profiles would be the autoregressive-moving average model (ARMA) and the autoregressive integrated moving average model (ARIMA). These techniques are presented in the literature as one of the best approaches for modeling univariate time series, which validates the selection made. It should be mentioned that both models have the same theoretical approach, the difference being that ARMA is used for stationary series and ARIMA for non-stationary series.

It has been mentioned repeatedly that in order to exploit to the maximum the usefulness of the reinforcement variable obtained from this stage, modeling and prediction must be the best possible for each user of the system. Therefore, given that an individual statistical model (ARMA or ARIMA) will be obtained for each client, for each of them the values of the degrees of the autoregressive ($\boldsymbol{P}$) and moving average ($\boldsymbol{Q}$) polynomials that allow the better fit of the consumption profile will be found. This is done using the Akaike Information Criterion (AIC), which is a method that allows the selection of the best ARMA/ARIMA model based on its performance and complexity.

The results obtained from the application of these models for the characterization of consumption profiles far outperformed those obtained initially with the intelligent algorithms. An example of the above is shown in Figure 5.3-1, which shows a consumption profile of a user of the system and its corresponding representation by obtaining the best ARMA/ARIMA model for that profile. It is possible to observe that the model is able to explain satisfactorily the consumption behavior of the client since it follows faithfully the original curve.

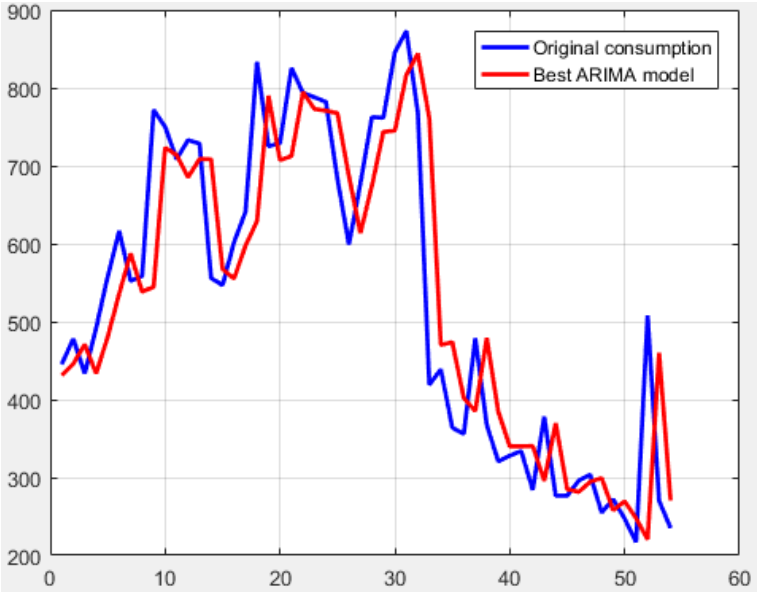


Fig 5.3-1. Example of ARMA/ARIMA adjustment of a consumption profile.

From figure 5.3-1 it can be seen that even the best model obtained is not able to explain perfectly the behavior of the consumption profile. This is because ARMA and ARIMA are univariate models of time series, that is, they try to explain the series with a single input variable, which is the same series. The difference between the original curve and that obtained with the ARMA/ARIMA model is a product of the inability of the original series to fully explain itself. In the particular case of the consumption of electrical energy there are factors that modify or alter the consumption behaviors of the users, these factors are known as exogenous variables. These exogenous variables are responsible for the fact that the consumption profile is not able to explain itself, and therefore, are the ones that allow to explain the differences between both curves.

5.3.3.2 Part 2: Intelligent correction of the statistical model

As it is desired to obtain the best possible model for each user, it is not enough to obtain the best statistical model (ARMA/ARIMA), since it does not explain the total consumption behavior. It was then proposed a complement that allows explaining the differences between the series of original consumption and its corresponding model that are product of the exogenous variables. Up to this point, of the input variables of this stage only the consumption profile has been used to obtain the

statistical model. The remaining inputs correspond to the group of exogenous variables that will be used to complement and improve the model by explaining the differences.

If obtaining the statistical model is the first part of this stage, the second part corresponds to the model that is responsible for explaining the differences between the original consumption curve and the statistical model that are the product of the exogenous variables. The idea is that at the end of the second part, the differences obtained are added to the statistical model to ***correct*** it and obtain a better representation of the user's consumption profile. Like the first part of the stage, the second part will also be executed individually for each user.

The example presented in figure 5.3-2 allows a better understanding of the concept of the differences between the actual series and the fitted model. The figure shows a real series defined by the observed values (***blue*** circles) and a model is fitted to represent (explain) the real series (***blue*** line). For each observed value there is a representation of the same by the model. It is defined as ***difference*** the distance between each actual observation and its corresponding representation in the fitted model. The ***differences*** (dotted ***red*** lines) are nothing more than what must be added at each point of the fitted model to perfectly explain the actual series of observations.

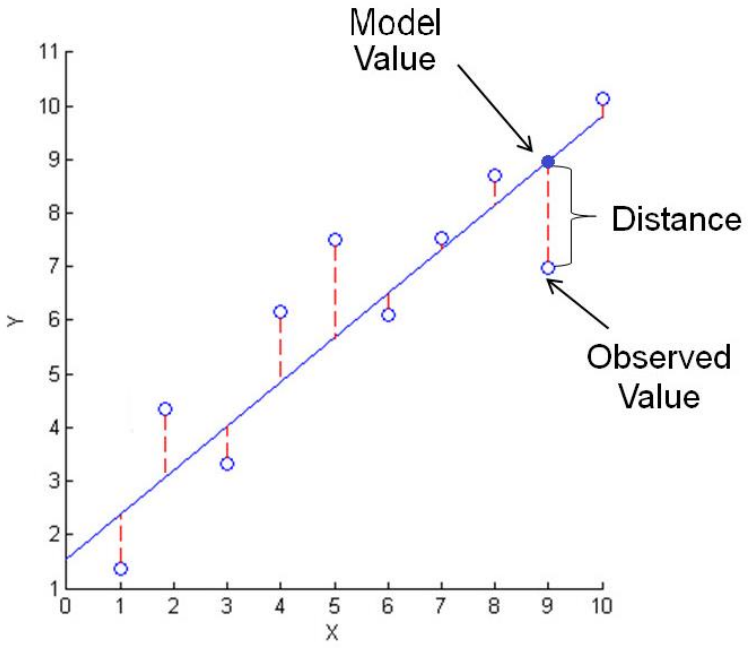


Fig 5.3-2. Example of explanation of the concept of ***difference***. (www.mathworks.com)

The obtaining of the ***differences*** corresponds to a problem of regression, which without having an analytical model of solution favors the use of computational intelligence techniques for this purpose. The intelligence technique selected must belong to supervised learning, this because all the ***differences*** are known (55 periods), which in this case represent the output. While the inputs are the variables ***Average temperature of the month***, ***Climate factor of the month***, ***User tariff*** and ***Days billed of the month*** for each of the 55 periods of the study window, which correspond to the exogenous variables of the model. From the above it is established that the set of examples that allow to train an intelligent algorithm in a supervised way are complete.

Currently, the two most widely used techniques for regression problems are neural networks and support vector machines. The technique selected for the second part of this stage was artificial neural networks in its configuration to solve this type of problems, which is a feed-forward multi-layer perceptron with training based on backpropagation. SVM works by solving an optimization problem, so for the same initial conditions it converges to nearby solutions. On the contrary, the neural network is initialized randomly before training, which allows to obtain solutions that can be completely different even with the same initial conditions of the data. This variability factor allows the exploration of an extensive set of solutions, opening the possibility of finding networks that generalize quite well and give way to obtaining better solutions. The above was the criterion for the selection of ANN as the technique to be used in this part of stage 2. To obtain the ***differences*** a set of several neural networks independent of each other will be trained, and the one that has presented the best performance will be selected.

At the output of the second part of this stage the set of ***differences*** that must be added to the statistical model in each period to perform the correction are obtained. Figure 5.3-3 shows an example of the intelligent correction process. It is possible to observe that there is a consumption profile of a user (***blue*** line), which is explained by the best statistical model found in the first part of the stage (***red*** line). It can be seen that the performance of the model is poor and does not allow

to characterize the behavior of the user. The execution of the second part of the stage gives the ***differences*** obtained from the network with better performance, which are added to the statistical model to obtain the model corrected from the second graph (***magenta*** line). It is clear that the intelligently corrected model succeeds in explaining satisfactorily the behavior of the user profile.

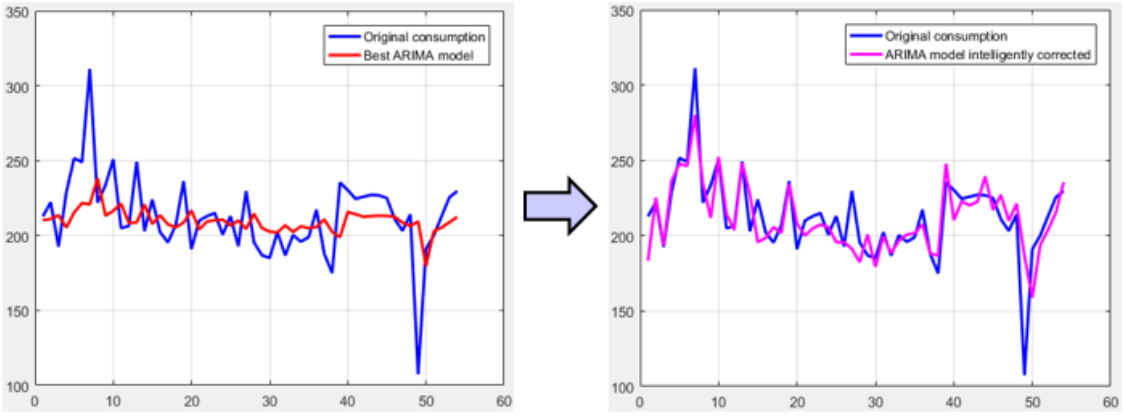


Fig 5.3-3. Intelligent correction of the statistical model by adding the ***differences***.

5.3.3.3 Selecting the prediction

After the first part, the ARMA/ARIMA model obtained is used to make the prediction of the next month's consumption, that is, the next value of the consumption series of the customer (consumption profile). Additionally, after the execution of the second part of this stage the ***differences*** for each value of the model are obtained, this includes the realized prediction. This indicates that the prediction is also corrected intelligently with its respective ***difference***.

Due to there is no guarantee that for all users is satisfied that smart correction improves the performance of the statistical model, there is no way to state that the most accurate prediction is the corrected one. Thus, the algorithm holds both the value of the statistical prediction and that of the corrected one. The prediction assumed to be valid is the one that is closer to the user's actual consumption value in the predicted month (when it occurs), as a way of not penalizing the user; since selecting the most distant one would induce an indicator of suspicious behavior for detection.

Figure 5.3-4 illustrates the complete execution of stage 2 of the system for a user, a zoom has been made over the last periods to better visualize the phenomenon. The consumption profile (***blue*** line) consists of fifty-four periods, the statistically and intelligently corrected models (***red*** and ***magenta*** lines) have fifty-five because they include the prediction. When the real value (dotted ***blue*** line) occurs, the comparison of both predictions with that value is made, the closest prediction in this example is the one of the intelligently corrected model. Therefore, the output of stage 2 of the system for said user would be the difference between the corrected prediction and the actual value, which has been defined as ***consumption deviation.***

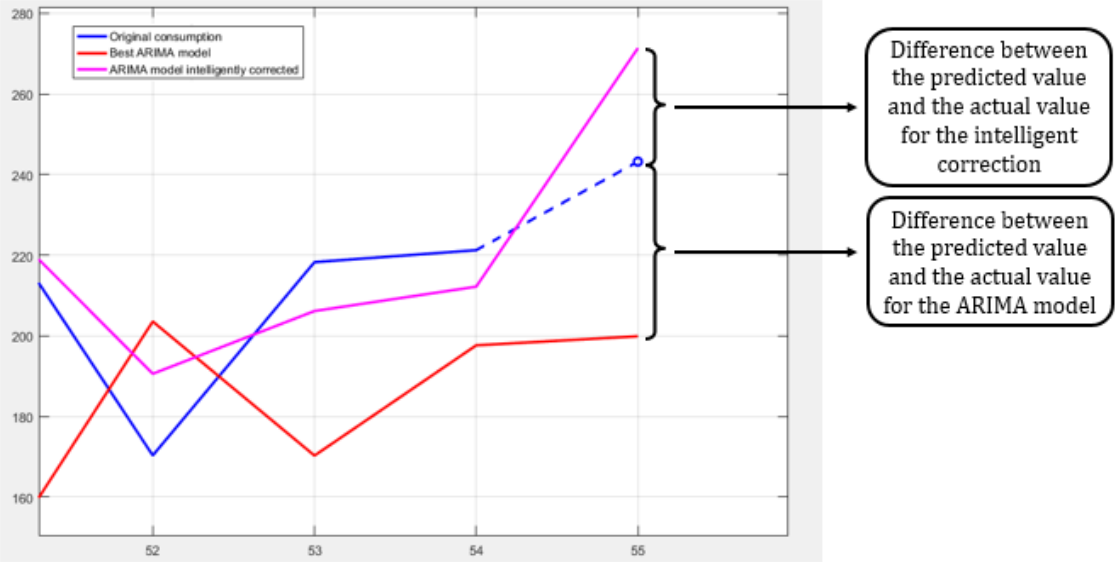


Fig 5.3-4. Example of the process of selecting the prediction for a user.

5.3.3.4 Functional structure of the stage

The previous approach allows defining the functional structure of the stage of consumption profile modeling and prediction of consumption of the next month. This stage consists of two parts, the first is the obtaining of the best model ARMA/ARIMA for each of the users of the system. These models are obtained by using the Akaike Information Criterion (AIC), which allows the selection of ***P*** and ***Q*** values that yield the best model in terms of performance and complexity. Once the models that represent each user are obtained, the predictions of the consumption of the next month are made for each one of them.

The second part uses neural networks to intelligently correct each of the user models obtained from the first part. This is done by selecting the best neural network to explain from the exogenous variables the ***differences*** in each of the statistical models of the users. The prediction of the next consumption is also affected by the correction.

Finally, statistically and intelligently corrected predictions are compared to determine which is accepted as valid for each user of the system. Selecting the one that is closest to the actual value of the customer profile to calculate the ***deviation of the consumption***. The functional structure of the stage is shown in Figure 5.3-5.

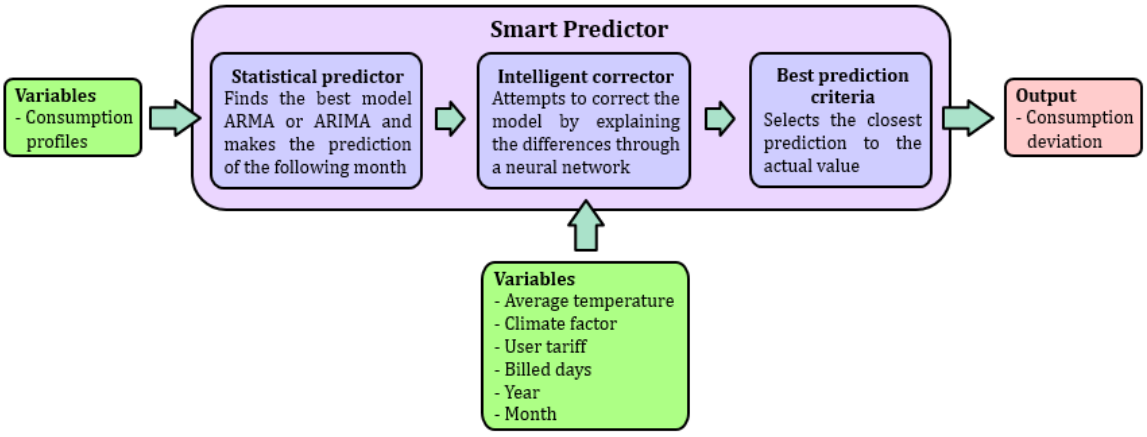


Fig 5.3-5. Functional structure of stage 2 of the system.

5.3.4. Conditions of execution

For the execution of the first part of the stage, the consumption profiles of all users of the system (92,794 users) are used. This is because a model must be obtained for each user, which indicates that each should be analyzed independently.

In the second part the percentage used for training is 60%, for validation is 20% and for testing is 20%. These percentages do not divide the set of users since each user is analyzed individually. These percentages divide the periods of a user's profile, that is, of the months that compose the customer consumption series, 60% are used to train the network, 20% to validate it and the remaining 20% to test it. This is the same for all the users of the system.

5.3.5. Implementation in Matlab

The implementation of this stage in the development environment will be described independently for each of its two parts.

5.3.5.1 Implementation of the obtaining of the statistical model

This section describes the aspects that were taken into account to obtain the statistical models of each user of the system.

- **Verification of stationarity:** To decide which model should be applied to characterize a user, it is necessary to check if the time series that results from the consumption profile of this is stationary or not. This is due to the existence of a specific model for each case. ARMA is suitable for stationary series and ARIMA for non-stationary series. The verification of the stationarity is carried out using the statistical test of Dickey - Fuller (Bohn, 2005). Which allows to extract the unitary roots of the series to verify the stationarity of the same.

  As a first step, the Dickey - Fuller test is applied to all consumption profiles of the users of the system. In this way it is defined which model should be applied to explain the behavior of each client.

- **Limits of the degrees P and Q of the model polynomials:** The lowest value that P and Q can take in ARMA/ARIMA models is zero. However, this value indicates the non-existence of the part of the model whose degree is zero. Since the fit of the best model for each user is desired, the existence of both the autoregressive part and the moving average part must be ensured. This indicates that the minimum value managed by the system for P and Q is one in both cases.

  The maximum value for these coefficients is given by the number of observations that exist from the time series to be modeled. The tests showed that the maximum value for P and Q from the 55 observations that compose the study window is two in both cases.

- **Selection of P and Q values for each user's model:** As mentioned in several times, the Akaike criterion allows to select the best model based on its performance and its complexity. For the ARMA/ARIMA models, the criterion measures the performance of the fitting by the likelihood of the model. To measure the complexity, the criterion quantifies the number of parameters of the model, this is represented by the sum of the degrees of $\boldsymbol{P}$ and $\boldsymbol{Q}$ of this. Let $\boldsymbol{LH}$ be the likelihood and $\boldsymbol{NumParam}$ the sum of $\boldsymbol{P}$ and $\boldsymbol{Q}$, the AIC expression to evaluate the models is:

$$AIC = -2(\boldsymbol{LH}) + 2(\boldsymbol{NumParam})$$

It was specified that for each user the models with degrees 1 and 2 would be tested for both $\boldsymbol{P}$ and $\boldsymbol{Q}$. So for each of these models their respective AIC is calculated, the model with the lowest AIC value is the best model. Table 5.3-1 exemplifies the AIC calculation for each of the possible models.

		Order of *P*	
		1	**2**
Order of *Q*	**1**	AIC(1,1)	AIC(2,1)
	2	AIC(1,2)	AIC(2,2)

Table 5.3-1. Calculation of the AIC for the possible models of a user.

5.3.5.2 Implementation of the intelligent correction

This section describes the aspects that were taken into account to perform the intelligent correction of the statistical models of each user of the system.

- **Number of hidden layer neurons:** This parameter is closely linked to the performance of the neural network. A low number of neurons will lead to poor performance, while a very high number can lead to overfitting. Experience has shown that the number of neurons in the hidden layer must be between the number of neurons in the input layer and the number of neurons in the output layer. To obtain this parameter there are several

heuristics according to the empirical rule mentioned above, the most common one being the following (Matich, 2001).

$$N_{Hidden} = \left[\left(\frac{2}{3}\right) * N_{Input}\right] + N_{Output}$$

Where N_{Input} is the number of neurons in the input layer, which corresponds to the number of input variables, six in the case of the second part of the stage. And N_{Output} is the number of neurons of the output layer, which is given by the number of outputs of the network. As for each case of the network execution, the values of the six inputs for a period allow to obtain as output the ***difference*** in said period of the real series and the statistical model; there is only one exit. Therefore, N_{Output} is one for the case of the second part of this stage.

The evaluation of the expression shows that N_{Hidden} is five for all networks to use within the second part of stage 2 of the system.

- **Number of neural networks:** It was mentioned earlier that the use of an ANN introduced a variability in the results, even though it would be executed with a fixed configuration. This happens when the problem is not easy to solve, allowing the network to learn differently in each execution. In order to take advantage of this feature, fifty independent networks will be trained for each user, the network with the best performance will be selected for the intelligent correction of the statistical model of that user.

5.3.5.3 Neural network performance metrics

For the evaluation of each of the networks trained to correct the model of a user was used a metric based on the linear correlation coefficient. It is well known that a way to evaluate the fit of a model is by calculating the correlation coefficient between the actual data and the fitted model data. The higher the value of the correlation coefficient the better the fitting performed by the model.

Since the data of the input variables is divided into three groups (training, validation and testing), the linear correlation coefficient for each group will be calculated. That is, compare the outputs of the network for each group with their

corresponding input data. Finally, the square of each of the three coefficients is computed and subsequently summed. The performance of a network is better the higher the calculated value.

$$Perf = {C_{Tra}}^2 + {C_{Val}}^2 + {C_{Test}}^2$$

Where C_{Tra} is the linear correlation coefficient of the fitting performed by the network for the training data, C_{Val} is the linear correlation coefficient of the fitting performed by the network for the validation data and C_{Test} is the linear correlation coefficient of the fitting performed by the network for test data. To obtain the definitive ***differences*** of a user the network whose ***Perf*** value is higher than the other forty-nine will be used.

5.4. Implementation of the stage of fraudulent users detection (stage 3)

As its name indicates, at this stage the detection of users with fraudulent connections takes place. All the efforts made for the implementation of the two previous stages converge at this point, this with the purpose of obtaining a detector with suitable performance that allows to contribute in the solution of the problem of non-technical losses that are the product of fraud. The function of this part of the system is to use the data of the variables it receives as inputs to generalize about the patterns and behaviors that differentiate a fraudulent user from a non-fraudulent one. For each user of the system, the output of this stage is a label that allows classification within one of two groups (fraudulent and non-fraudulent).

5.4.1. Input variables

The input variables used for the execution of this stage are mentioned below.

- User measurement equipment.
- Geographic location of the user.
- Number of complaints made in the last year.
- Number of overdue bills.
- Customer fraud history.

- Reading and billing anomalies in the last year.
- Group to which each user belongs (output of stage 1).
- Consumption deviation of the month (output of stage 2).

With the exception of the variables obtained as outputs in stages 1 and 2 of the system, all other input variables for this stage are obtained from the user database supplied by the commercialization company of the region.

Since the execution of the system to detect fraudulent users is done on a monthly basis, the variable ***Consumption deviation*** is composed of the deviations of users for the month in which the detection of frauds is desired. The variables ***Number of complaints made*** and ***Reading and billing anomalies*** are evaluated for the last twelve periods before the month of execution of the system, this because they represent user behaviors that can change with the passage of time. The variables ***User measurement equipment***, ***Geographic location*** and ***Group to which the user belongs*** are fixed. Finally, the variables ***Customer fraud history*** and ***Number of overdue bills*** result from evaluating the user throughout the study window.

5.4.2. Conditioning of input variables

In the previous chapter it was mentioned that the conditioning done to most of the input data of stage three of the system was a transformation to the binary representation of these. It was explained that for the classification problems it has been demonstrated that this transformation facilitates to the intelligent algorithm the recognition of patterns, increasing its capacity of generalization and therefore, improving its performance.

The transformation mentioned above was performed on all input variables that are integer numerical, which in this case are all except the ***deviation of consumption***, which remains the same. The process of binary representation of inputs will be explained by the example in Figure 5.4-1.

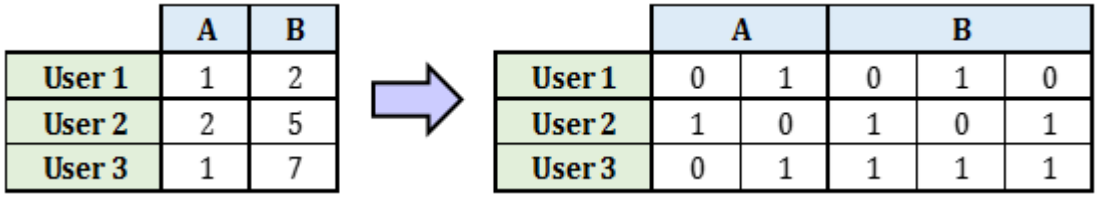


Fig 5.4-1. Example of the binary representation of the inputs.

Suppose that two variables ***A*** and ***B*** to characterize a set of three users of any process are given. The first table of the figure shows the values of the two variables for each user, where it is possible to observe that all are integers. The binary representation process finds the highest value of each of the two variables, being 2 for ***A*** and 7 for ***B***. The above since the largest number of a variable indicates the maximum number of digits to be used for the binary representation of the same. The number 2 in binary is 10, so that two digits are used for variable ***A***; while the number 7 in binary is 111, so that three digits will be used for the variable ***B***. After this, the binary conversion of all the values of each variable is performed according to their respective maximum numbers of digits, from which the second table is obtained. This table is the one used as new input data to the intelligent algorithm. It can be observed that the number of inputs increased from two to five, however, a better performance of the technique is expected with the use of this transformed data.

5.4.3. Structure of the stage

5.4.3.1 Obtaining fraudulent and non-fraudulent users for training

In the previous chapter, the conceptual design of the stage of detection of fraudulent users was presented (see chapter 4, section 4.2.2.4), which functionally corresponds to the development of a supervised classification task. The above because before being used as a fraud detection tool, the detector must be trained to recognize them by a set of examples with known inputs and outputs. In this way it is given the ability to recognize the patterns and generalize the behaviors that allow to define if a user is fraudulent or not.

The dataset containing the input variables of this stage has information from all the clients of the system, that is, the variables are available for both types of users (fraudulent and non-fraudulent). However, because the training is supervised, for each user the output defined by the input variables must be known. In other words, one must know if a user is fraudulent or not based on the values of their inputs.

In section 4.2.1, some of the system variables of the type "***Classification of user type***" were proposed, which represent the proposal of this investigation to obtain the output set that allows the training of the detector. The above because although the values of the inputs for each user are known, there is no direct way of knowing if they are fraudulent or not and the supervised training of stage 3 could not occur. The use of the variables of this type to obtain the set of outputs for training is explained below.

Non-fraudulent users

To obtain the examples of training of non-fraudulent users, the following variables were used:

- Balance of macrometers.
- Moorage of macrometers.

It was explained that a macrometer is a measurement equipment located at the output of the distribution transformers, and that the users connected to said transformers each have their respective measurement equipment. ***Balance of macrometers*** contains the information of the consumption of a macrometer, while ***Moorage of macrometers*** contains the information of the consumptions of the meters derived from said macrometer. It was defined that non-fraudulent users would be accepted as those where the difference between the consumption of the macrometer and the sum of the consumptions of the derived meters does not exceed 5%, which is a value given entirely by technical losses. From this criterion the outputs for the examples of non-fraudulent users that will be used in the training of the detector are obtained.

Fraudulent users

Contrary to the previous case, there is no criterion to assume if a user is fraudulent. However, it is known that the commercialization company performs a set of monthly user inspection visits. Therefore, the only possible way to recognize fraudulent users is by reviewing the results of the inspections carried out during the monthly energy recovery campaigns. These results are reported in the variable ***Results of campaigns and inspections***, which indicates whether an inspected user was found to have committed fraud or not.

Using the above variable, users who have been found with fraudulent connections will automatically be included in the base of examples of fraudulent users that will be used in the training of the detector.

5.4.3.2 Supervised Classification Algorithm

When the properly conditioned inputs and the outputs corresponding to these inputs are entered, it is possible to continue with the training and subsequent execution of the intelligent classification algorithm. There are many computational intelligence techniques oriented to the solution of the supervised classification problem, however, tests were only performed on the three most widely used. These are support vector machines, artificial neural networks and random forests. Unlike the previous two stages of the system, where there were specific criteria for the selection of the technique to be used; for this stage all the options are equally valid. The selection of the technique took place after the tests, where the one with the best performance in the detection of fraudulent users was chosen.

5.4.3.3 Functional structure of the stage

The previous approach allows defining the functional structure of this stage, which performs a conditioning of most of its input variables based on a binary representation of them, leaving the others intact. Specially selected variables are used to obtain the set of outputs for the training. The computational intelligence technique used for supervised user classification is random forests. The output is a label for each user where the class is indicated, with the possible options being

fraudulent and non-fraudulent. The functional structure of the stage is presented in Figure 5.4-2. Arrows with dotted lines indicate that this part of the process occurs only during training, and is not taken into account during the commissioning of the system.

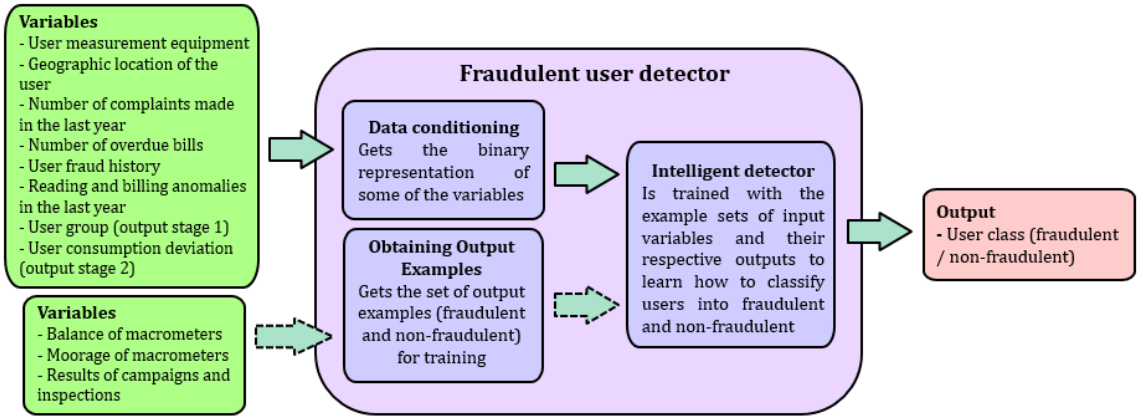


Fig 5.4-2. Functional structure of stage 3 of the system.

5.4.4. Conditions of execution

For the execution of the stage, the division of the dataset for the training and testing was not performed on the number of users but on the number of periods of the time window. The study window of the system is of 55 periods, starting in January 2011 and ending in July 2015. For functional requirements, the periods from January 2011 to October 2014 are reserved for the execution of stage 2 of the system, due to the fact that a lower number of periods significantly reduces the performance of the intelligent correction. The periods from November 2014 to April 2015 are used for the training of the detector and finally, the periods from May 2015 to July 2015 have been used to test the system. The reason behind all this approach is the monthly execution feature of the system.

As a result of this, information about a user must be kept for a minimum of three years and ten months (46 periods) before being analyzed by the proposed system. This condition is imposed by the second part of stage 2 (intelligent correction), which requires at least this time to achieve adequate performance. It is worth noting that the system can be executed for a lower number of periods, however, this should be done using only the first part of stage 2. Stage 1 (cluster) can be run

with any number of periods, while stage 3 (detector) requires at least 6 training periods to achieve suitable performance, which have been guaranteed in the division of the window proposed in the first paragraph of this section. It should be noted that everything mentioned in this section is a product of tests performed during the design and implementation of each stage of the system.

5.4.5. Implementation in Matlab

Next, the relevant aspects that were taken into account for the implementation of this stage in the development environment are presented.

5.4.5.1 Implementation of the classifier based on random forests

- **Number of trees in the forest:** Values of the number of trees were tested from 10 to 100. Above this value the stage execution time increased considerably without a noticeable increase in performance.
- **Number of repetitions:** 10 executions are performed for each value of the number of trees.

5.4.5.2 Classifier performance metrics

The most commonly used metric to evaluate the performance of classification algorithms is the confusion matrix (Fawcett, 2005). Therefore, it will be the one that will be used in the evaluation of the performance of the detector (and the system). The Figure 5.4-3 shows the structure of a confusion matrix to evaluate the solution of a ***binary classification***[2] problem. The columns represent the actual classes of the elements and the rows represent the class in which the algorithm placed them. The ***green*** cells indicate the elements that the algorithm correctly located in its real class, while the ***pastel orange*** cells indicate the elements that the algorithm did not locate in the class to which they actually belonged. The performance of the classification is obtained by calculating the percentage of correctly classified items (***green*** cells) with respect to the total of the same.

[2] A ***binary classification*** problem is defined as the one in which the elements to be classified can only belong to one of two possible classes. The problem of fraud detection is a binary classification problem since it is desired to classify users into fraudulent and non-fraudulent.

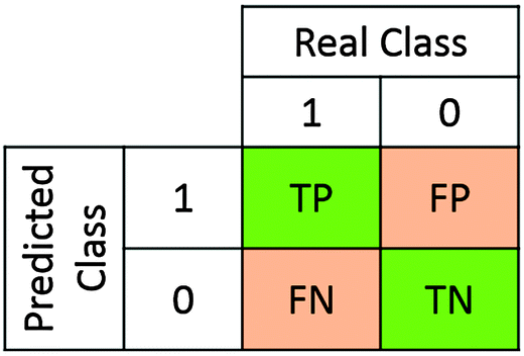


Fig 5.4-3. Confusion matrix to evaluate a binary classifier. (www.pubs.rsc.org)

PART III

EXPERIMENTS RESULTS AND CONCLUSIONS

Chapter 6

Analysis of the Experimental Results

This chapter presents the discussion and analysis of the empirical experiments and testing that have been carried out for the proposed test beds. The results depicted in this chapter demonstrate the utility, feasibility and reliability of the overall proposed approach presented in the previous chapters.

In the previous chapters, the design and implementation of the intelligent system for detecting fraudulent users proposed in this research were presented. Before being used as a fraud detection tool, the system must undergo an evaluation and testing process to validate its functionality. This is the purpose of this chapter, in which the tests that were performed are presented and the results obtained are analyzed.

Since the proposed system consists of three stages with different objectives, it became necessary to independently evaluate and validate each of these. Therefore, the chapter is divided into three large sections, each designed to present the results of a specific stage. The evaluation of stage 3 corresponds to the evaluation of the functionality of the system, since it is in this stage where the classification that allows the detection of fraudulent users is carried out.

6.1. Tests and results of stage 1: Grouping of consumption profiles

The hybrid design between the clustering technique and the genetic algorithm was proposed due to the need to find the best possible grouping. The function of the genetic algorithm is to explore the solution space composed of possible combinations of the SOM configuration parameters. The use of GA allows to reach the best solution without the need to explore all the points of the solution space, this due to its particular "heuristic" of operation. This avoids the need to use a design of experiments, mainly because the execution of the GA guarantees the convergence towards the best solution within the feasible space.

Next, the tests performed for the establishment of the upper limit of the dimension parameter are presented, as well as the results of the execution of the stage.

6.1.1. Selecting the upper limit of the SOM *dimension* parameter

It was mentioned in Sections 5.2.5.1 and 5.2.5.2 that the selection of the upper limit for the SOM ***dimension*** parameter had been established experimentally. Likewise, it was explained that a criterion was established to increase the performance of the stage without falling into ***individualization***. This criterion consists in finding the groups with greater and smaller number of elements after the grouping, if the number of elements in the minor group does not exceed 5% of the number of elements of the major, the grouping is considered invalid.

The tests performed consisted of the successive execution of the SOM by varying the upper limit of the dimension parameter. It was initiated from the value of the lower limit that is 2 (4 groups), the grouping was carried out and the established criterion was applied to evaluate the results. It should be noted that the same configuration was tested a total of ten times to take into account the variance of the process. If the results obtained from the configuration under study met the criterion, the value of the upper limit was increased by one and the process was carried out again.

The tested upper limit values met the criteria with no problem in all of their replicates until the number 15 (225 groups). In number 16 (256 groups), of the 10 tests performed 5 met the criteria and the other 5 did not. Due to lack of certainty to make a decision, the test with the number 17 (289 groups) was carried out. The results showed that in none of the 10 SOM executions the results were accepted. Therefore, it was decided to set the value of the upper limit in 16 (256 groups), in order to include those results that fulfill the condition for this value.

6.1.2. Pilot test of stage 1, set of 1000 users

It is known that the total number of users of the system is 92,794, however, a pilot test was designed with 1000 users to explain the structure of the results that are obtained when executing the stage.

6.1.2.1 Optimization results

The execution of stage 1 of the system concludes when the optimization process is finished. It should be remembered that the optimization performed is multi-objective, in which it seeks to obtain the least number of groups and the best performance. The multi-objective optimization processes do not yield a single solution, they yield a set that contains the best solutions for each objective function. The above because not necessarily a single solution meets the multiple objectives pursued. The set containing the best solutions for each objective function is known as the Pareto frontier.

The completion of the GA yields the Pareto frontier with the best solutions to the problem. For the purpose of system automation, the algorithm chooses by default the solution with the best value of the proposed performance metric, implying that the goal of the number of groups is sacrificed, which leads to a greater number of these.

In executing the stage 1 for the pilot test, the Pareto frontier shown in Figure 6.1-1 was obtained. It should be noted that the application of the criterion to avoid ***individualization*** in this pilot test showed that the upper limit of the ***dimension*** parameter is 8 (64 groups). It is possible to observe that the solution with the lowest value of the performance metric (best value) leads to the largest number of groups, whereas the solution with the smallest number of groups leads to the highest value of the performance metric. This example allows a clear understanding of the Pareto frontier concept and the best solutions for each objective.

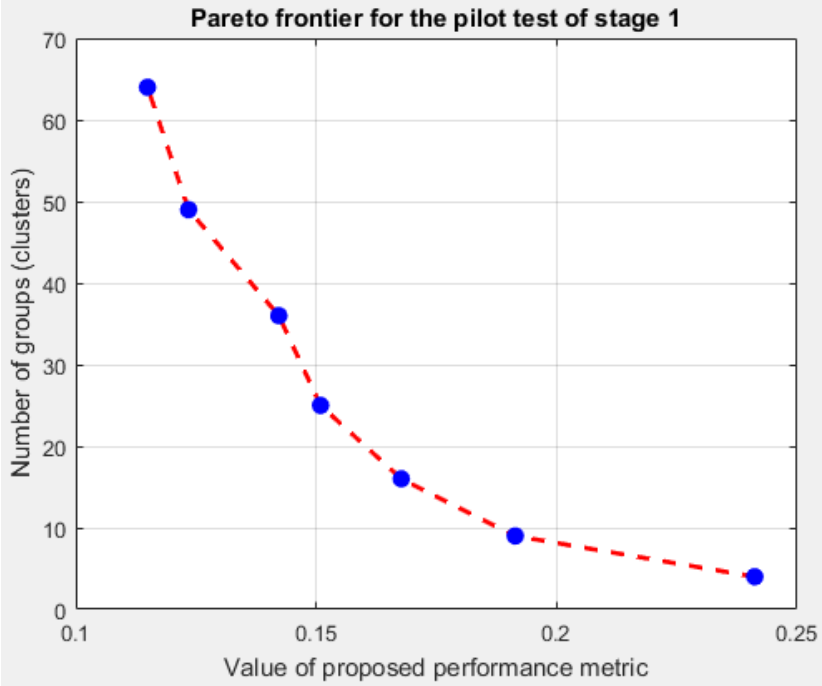


Fig 6.1-1. Pareto frontier of the pilot test with 1000 users.

Table 6.1-1 contains the values of the two objective functions that compose the Pareto frontier.

Values of the objective functions at the Pareto frontier	
N° of clusters	**Value of performance metric**
4	0,2412
9	0,1914
16	0,1677
25	0,1509
36	0,1422
49	0,1235
64	0,1149

Table 6.1-1. Pareto frontier values.

It was mentioned that the algorithm assumes by default the solution with the best value of the proposed performance metric, which for this case corresponds to the point of the frontier that leads to obtaining sixty-four groups and the lowest value

of the metric (0.1149). The values of the SOM parameters that describe the configuration representing that frontier point are:

- **Dimension:** 8 x 8 neurons (64 groups).
- **Topology:** Grid.
- **Distance function:** 'linkdist'.
- **Number of training steps:** 100 steps.

6.1.2.2 Results of the grouping

The grouping shown in Figure 6.1-2 is obtained from the execution of the SOM with the configuration described above. Which shows the number of elements that result in each of the sixty-four clusters of the square network. The group with the highest number of elements contains 51 users, which indicates that the minimum number of users in a group must be 3, which is fulfilled in the grouping shown.

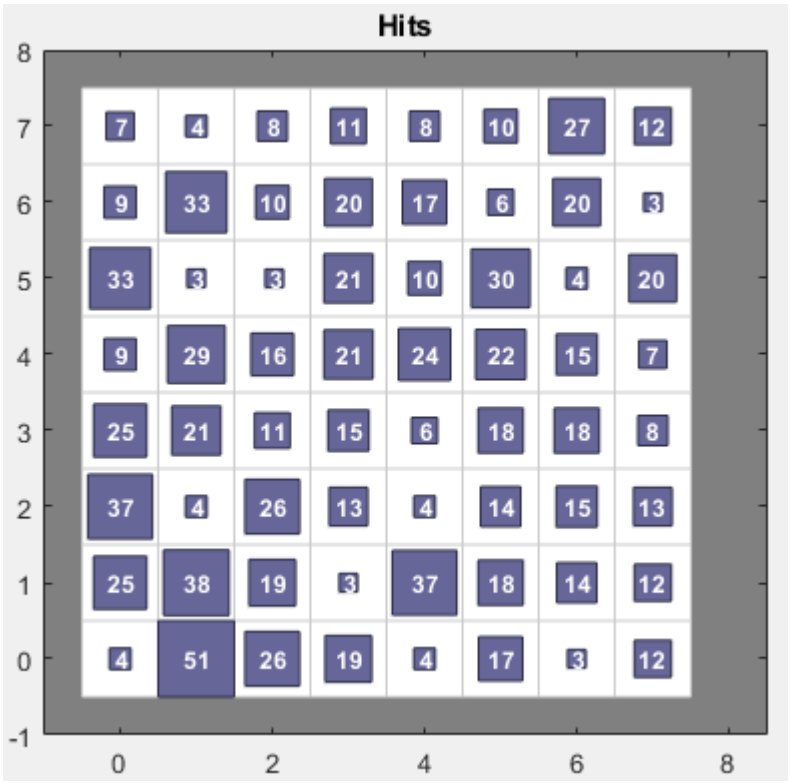


Fig 6.1-2. Number of users that result in each group after execution.

Next, the graphs of the resulting consumption profiles in each group after the clustering will be presented. However, due to the technical nature of this document

it is not possible to present the graphs of the sixty-four groups, therefore, a sample of eight of them has been selected for the inclusion in it.

The consumption profiles are composed of 55 periods and correspond to the curves with ***blue*** lines. The ***red*** line corresponds to the average profile of each group.

Clusters 4 and 5

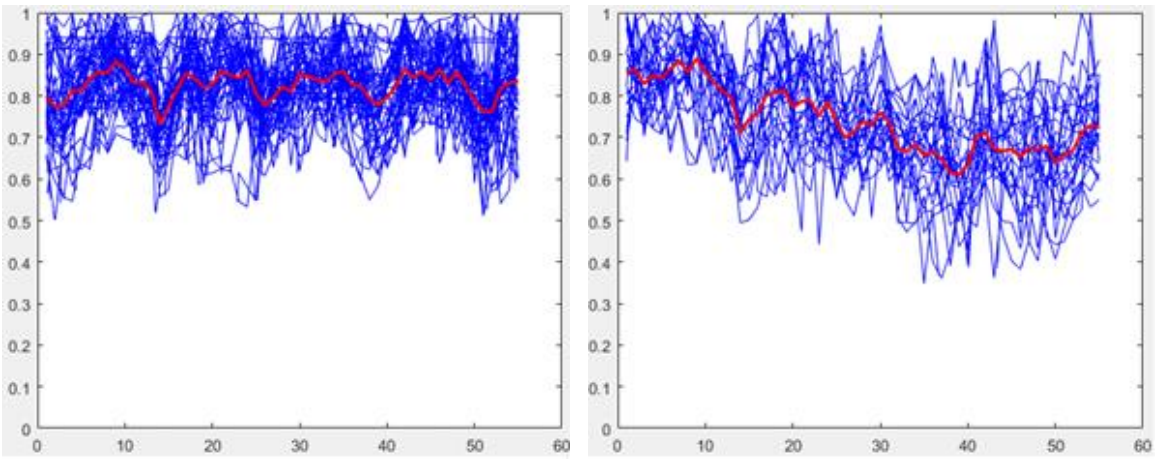


Fig 6.1-3. Consumption profiles of groups 4 and 5.

Clusters 8 and 9

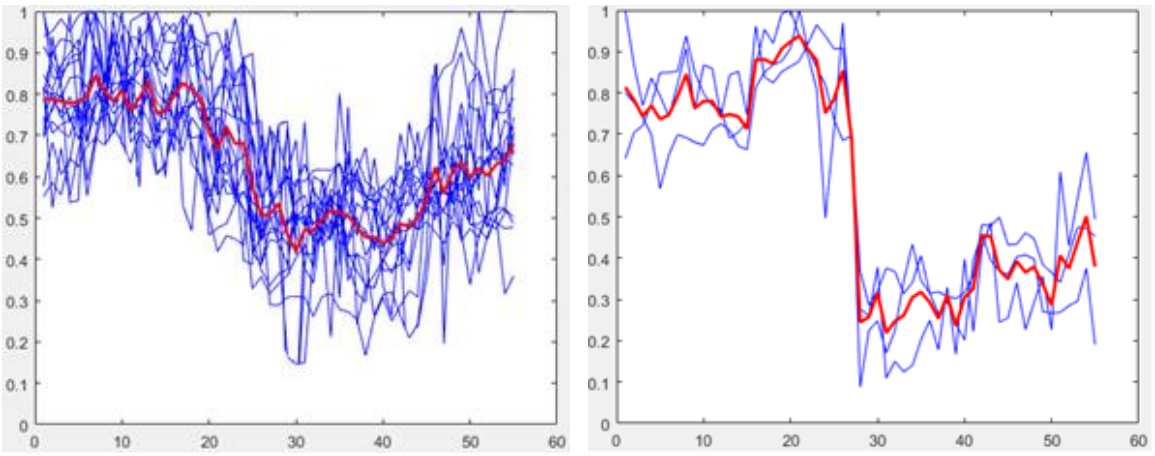


Fig 6.1-4. Consumption profiles of groups 8 and 9.

Clusters 14 and 19

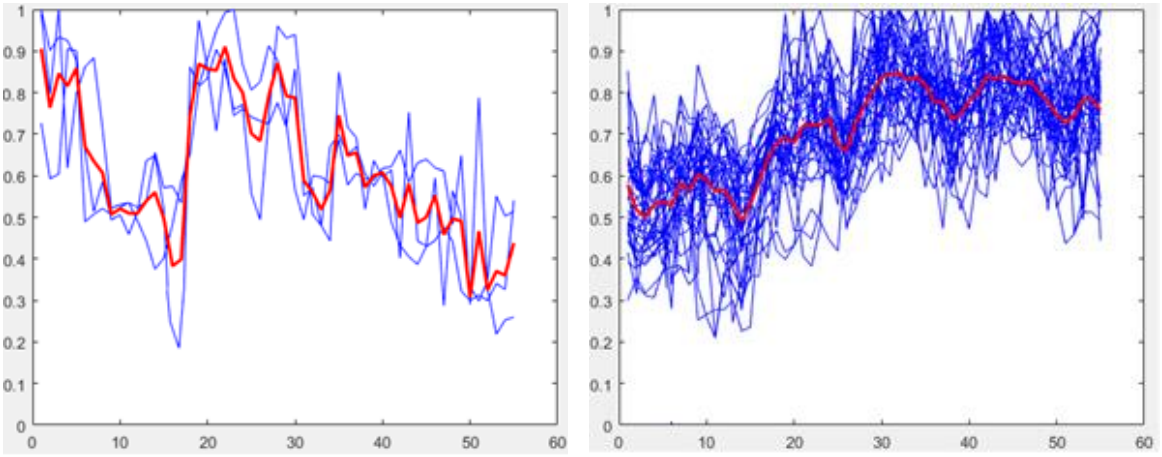


Fig 6.1-5. Consumption profiles of groups 14 and 19.

Clusters 34 and 62

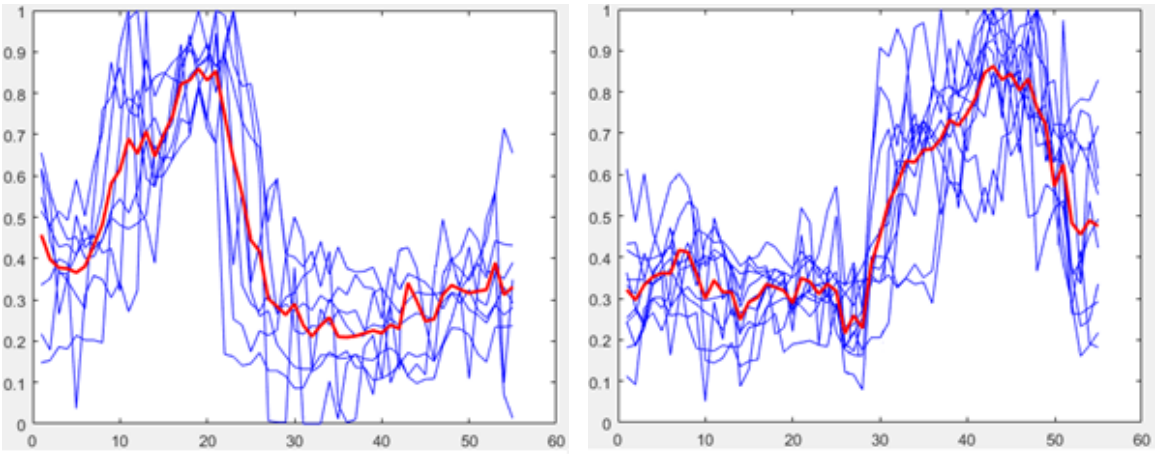


Fig 6.1-6. Consumption profiles of groups 34 and 62.

6.1.3. Definitive execution of stage 1, set of 92,794 users

After explaining the process of obtaining and analyzing the results of stage 1, the results obtained from the final execution of this stage are presented, which was performed with the total users of the system (92,794). It was mentioned in section 5.2.5.1 that the solution space consisted of 360 combinations of the SOM configuration parameters, the GA had to evaluate 217 of these combinations to reach at the solution set (Pareto frontier) of the optimization. This validates the decision to use this metaheuristic technique over an experimental design. The

Pareto frontier obtained after the execution of the stage for the total set of users is shown in Figure 6.1-7.

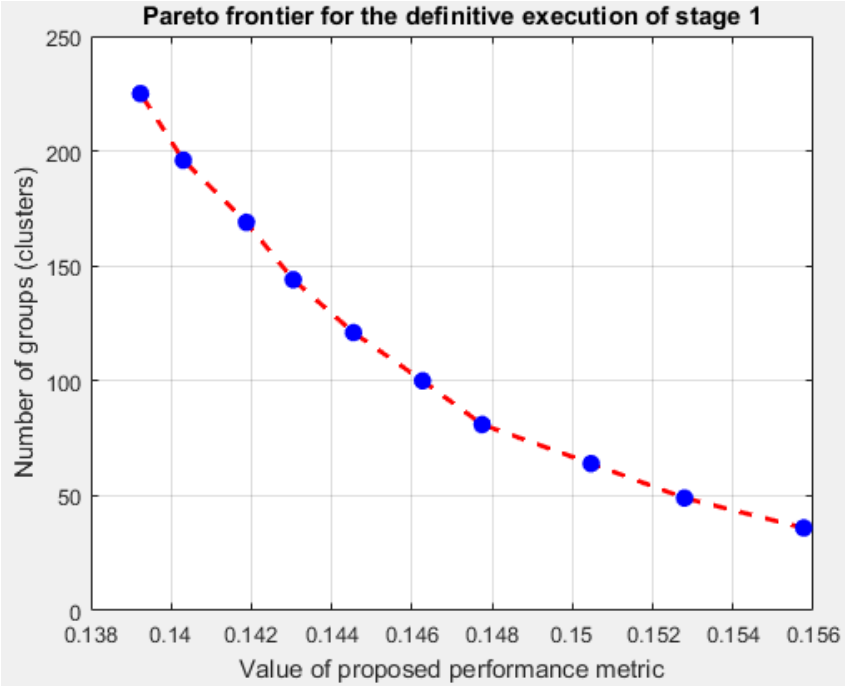


Fig 6.1-7. Pareto frontier of the definitive execution.

Table 6.1-2 contains the values of the two objective functions that compose the Pareto frontier.

Values of the objective functions at the Pareto frontier	
N° of clusters	**Value of performance metric**
36	0,1558
49	0,1528
64	0,1505
81	0,1477
100	0,1463
121	0,1445
144	0,1430
169	0,1419
196	0,1403
225	0,1392

Table 6.1-2. Pareto frontier values.

It is possible to observe that the Pareto frontier does not include the value of the upper limit of the parameter ***Dimension*** (16). This is because the "chromosomes" of the GA that included this value of the parameter, when evaluated did not meet the proposed criterion of the minimum number of elements in a group.

The default selection of the solution is made with the best value of the proposed performance metric, which corresponds to the point of the frontier that leads to the obtaining of two hundred and twenty-five groups (225) and the lowest value of the metric (0.1392). The values of the SOM parameters that describe the configuration representing that frontier point are:

- **Dimension:** 15 x 15 neurons (225 groups).
- **Topology:** Grid.
- **Distance function:** 'boxdist'.
- **Number of training steps:** 100 steps.

The grouping shown in Figure 6.1-8 is obtained from the execution of the SOM with the configuration described above. The group with the highest number of elements has 1237, while the group with the least elements has 140, this value is 11.1% of 1237. Therefore, it fulfills the criterion of minimum number of elements.

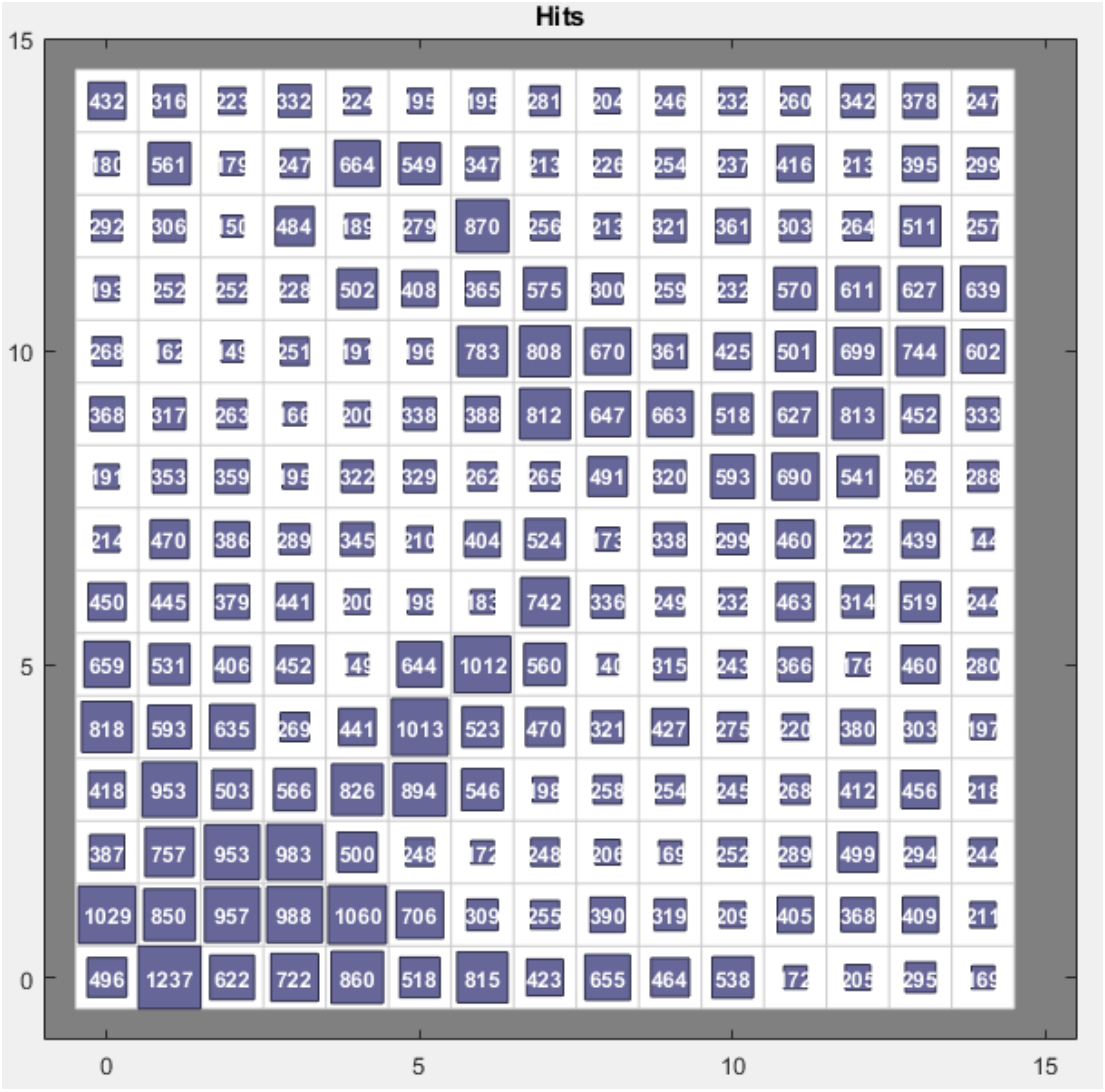


Fig 6.1-8. Number of users that result in each group after the clustering.

In order to present the results of the final execution of the SOM a case similar to the one of the pilot test occurs, since, due to the inability to show the graphs of the consumption profiles located in each one of the 225 groups, a sample of eight of these is presented.

For the graphs presented, the consumption profiles are composed of 55 periods and correspond to the curves with ***blue*** lines. The ***red*** line corresponds to the average profile of each group. The selected groups are shown below.

Clusters 1 and 31

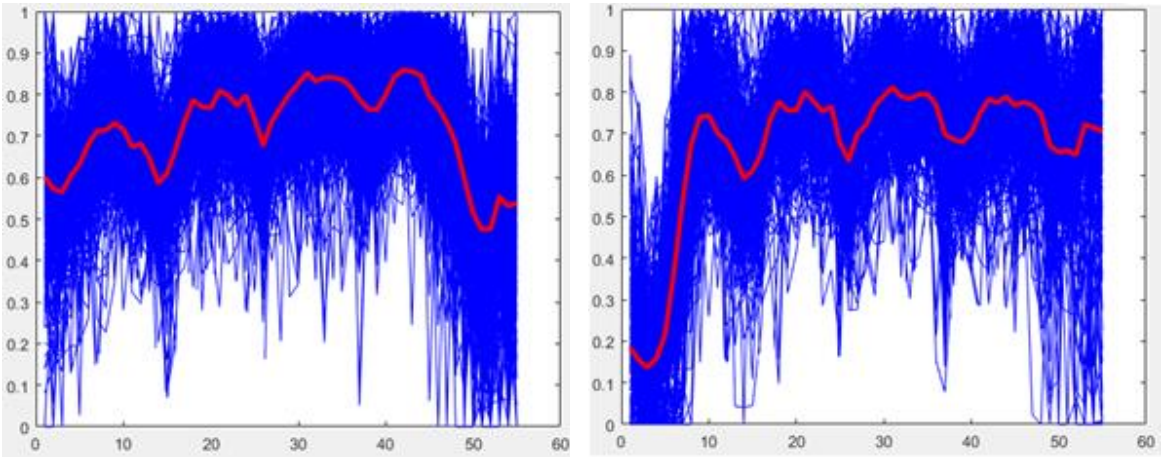


Fig 6.1-9. Consumption profiles of groups 1 and 31.

Clusters 18 and 162

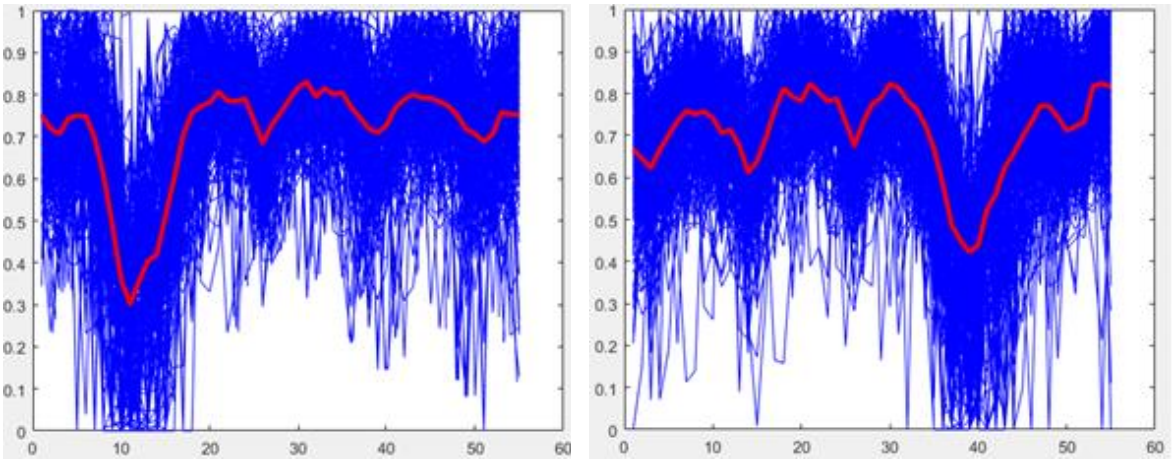


Fig 6.1-10. Consumption profiles of groups 18 and 162.

Clusters 117 and 130

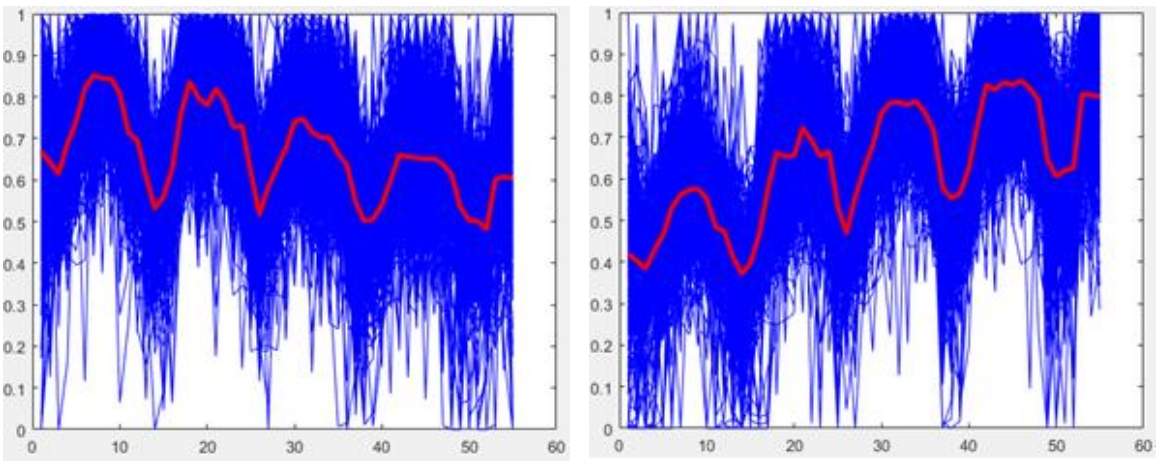


Fig 6.1-11. Consumption profiles of groups 117 and 130.

Clusters 57 and 97

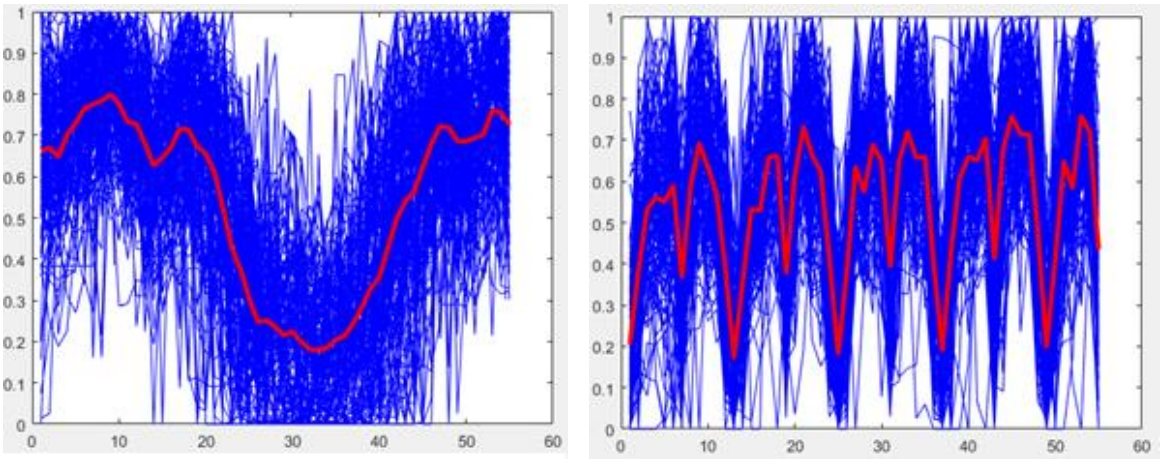


Fig 6.1-12. Consumption profiles of groups 57 and 97.

By performing a visual inspection of the examples presented for each grouping and reviewing the results and analysis of the Pareto frontiers, it is possible to conclude that the GA is able to find the configurations that make the best groupings according to the objective functions for each case. The default selection of the best performing solution allows obtaining the desired functionality, which is evidenced by observing that in each group, consumption profiles with similar behaviors (shapes of the curves) are located, thus validating the operation of this stage of the system.

6.2. Tests and results of stage 2: Modeling and prediction of consumption profiles

The previous chapter explained the process of implementing the stage 2 of the system, which is divided into two parts. The first one is a statistical modeling of consumption profiles, where an ARMA/ARIMA model is fitted for each user of the system, these models must correspond to the one that best explains the profile of each user. To achieve this, the Akaike Information Criterion is used, which allows comparing ARMA/ARIMA models of different orders and selecting the one that performs well with acceptable complexity.

In an attempt to improve the fitting made by the statistical models, the second part of the stage was implemented, which trains a set of fifty independent neural networks to select the one that is able to better explain the ***differences*** between the actual consumption and the fitted model. Like the first part, this second part is also executed individually for each of the users of the system.

Next, the tests performed on each of the parts of stage 2 are described. This to evaluate the performance and validate the functioning of these parts. Due to both parts are independent of each other, the tests and experiments, as well as the analysis of the results obtained are presented separately for each of these.

6.2.1 Obtaining the statistical model of each user

After validating the stationarity of the profile of each user by means of the Dickey - Fuller test, for each one of them, the corresponding four statistical models are fitted. This is due to the fact that the combinations (1,1), (1,2), (2,1) and (2,2) of the degrees ***P*** and ***Q*** of the polynomials of the model are tested. Subsequently, the best of the four ARMA/ARIMA models fitted is selected using the Akaike Information Criterion. Although the AIC allows to select the best model based on the likelihood and complexity of the model, the most commonly used metric to evaluate the performance of curve fitting models is the Mean Absolute Percentage Error (MAPE) (Adhikari & Agrawal, 2013).

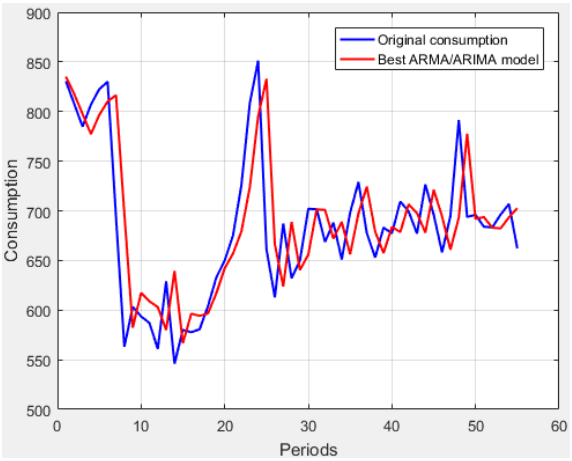
$$MAPE = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

Where ***n*** is the number of observations in the series, A_t is the actual value at time ***t*** and F_t is the value fitted by the model at time ***t***. This metric will be used to quantify the performances of the statistical models obtained for each user as a product of the first part of this stage. The above in order to have a metric to evaluate the performance of the statistical modeling from a general perspective, which is used to validate the operation of this first part of the stage as a whole.

As the execution of this first part is individual for each user, a total of 92,794 ARMA/ARIMA models were obtained. Due to the impossibility of showing them all

in this document, a selected sample of eight users is presented which allows to verify the functionality of this part of stage 2.

Selected user 1



Selected user 2

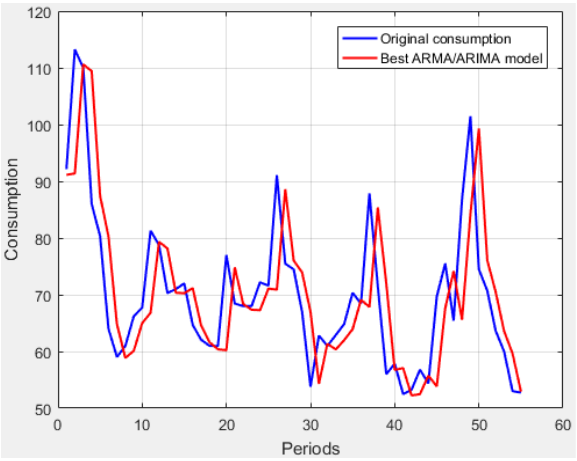
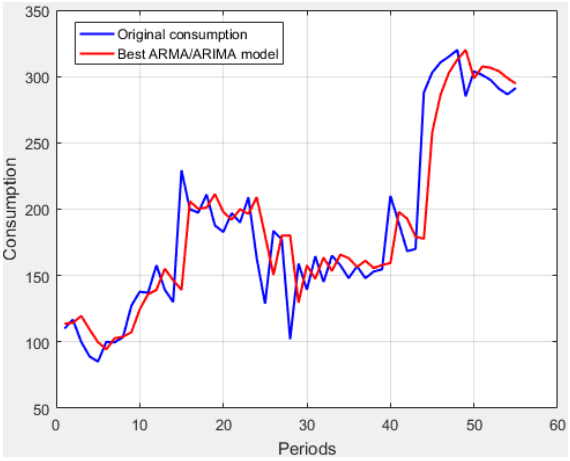


Fig 6.2-1. Best statistical models obtained for sample profiles 1 and 2.

Selected user 3



Selected user 4

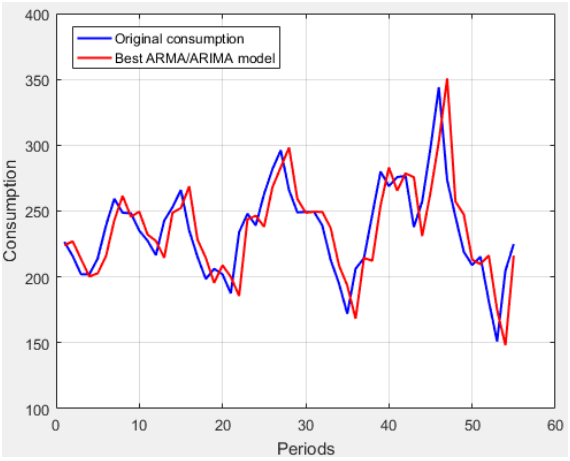


Fig 6.2-2. Best statistical models obtained for sample profiles 3 and 4.

Selected user 5 **Selected user 6**

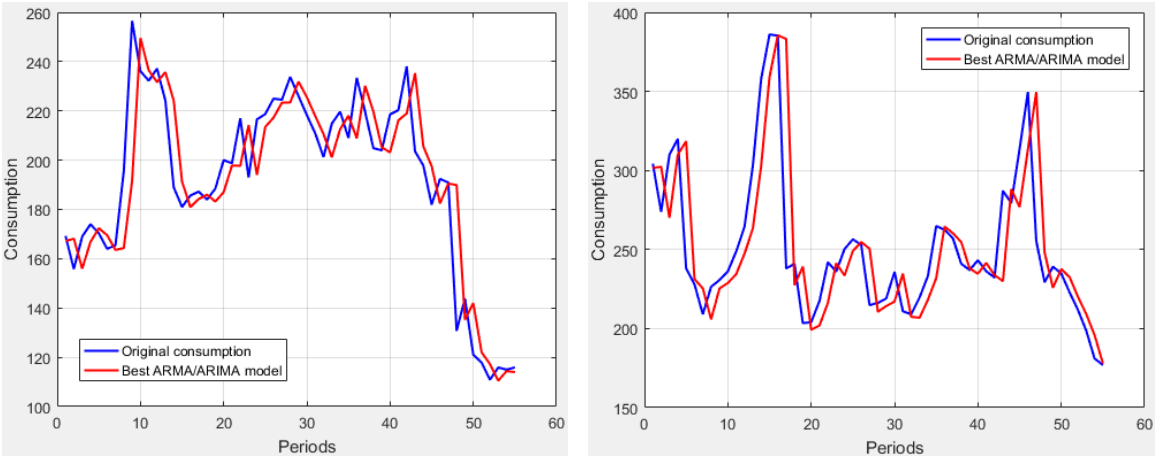


Fig 6.2-3. Best statistical models obtained for sample profiles 5 and 6.

Selected user 7 **Selected user 8**

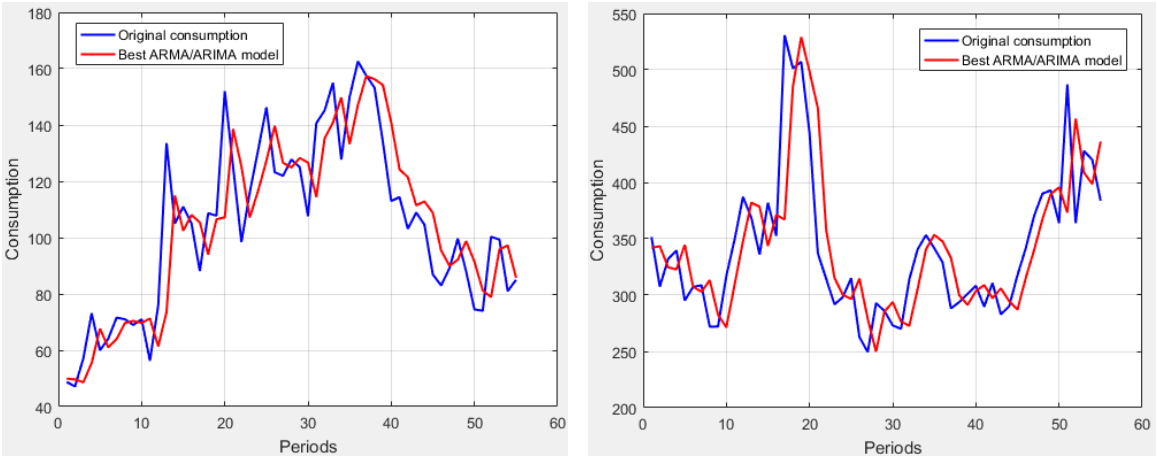


Fig 6.2-4. Best statistical models obtained for sample profiles 7 and 8.

Table 6.2-1 presents the MAPE for each of the eight examples of fittings presented above.

Performance of sample models	
Selected user N°	**MAPE (%)**
1	5,48
2	10,35
3	11,43
4	7,87
5	6,44
6	8,31
7	12,12
8	8,88

Table 6.2-1. MAPE for each model of the sample users.

6.2.2 Overall performance of the first part of the stage

In order to evaluate the overall performance of the fittings made by this part, the graph of Figure 6.2-5 is presented, which shows the distribution of all the MAPE obtained for the users of the system. The average of the MAPE of the 92,794 users is **15.42%,** which is illustrated in the graph with the horizontal ***red*** line.

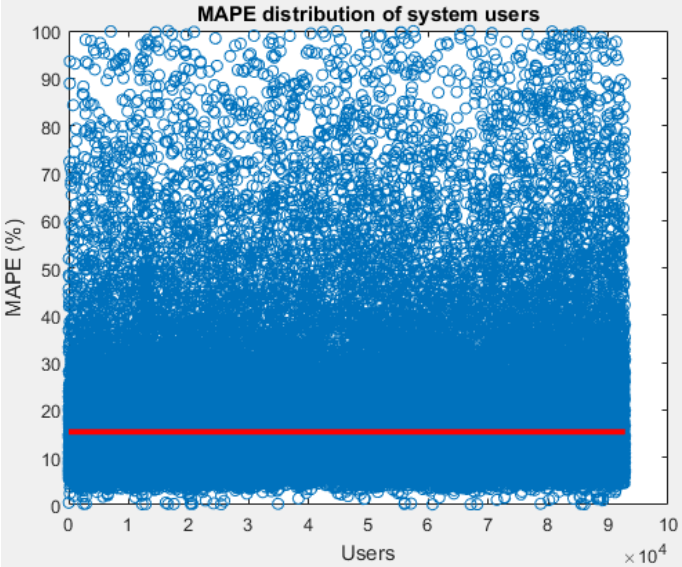


Fig 6.2-5. Distribution of MAPE for all users of the system.

In addition, the histogram corresponding to all the MAPE obtained from the first part for the complete set of system users is presented in figure 6.2-6.

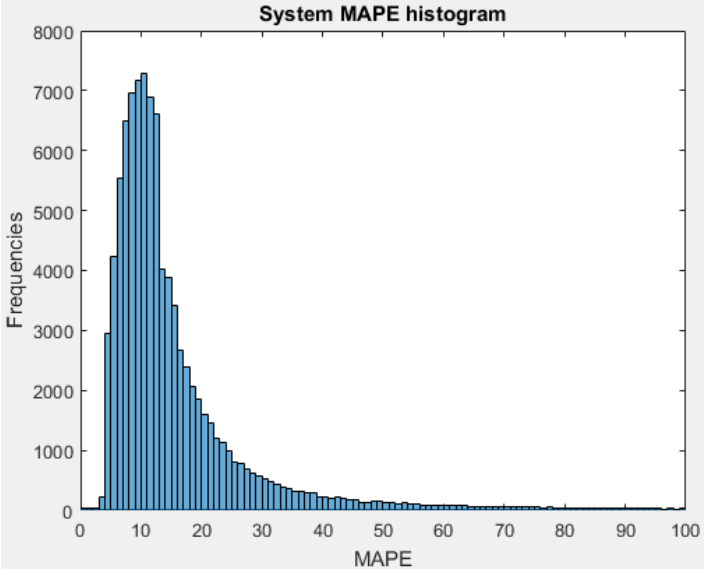


Fig 6.2-6. Histogram of the system MAPE.

The MAPE that were obtained from the selected sample of eight models of system users presented in Section 6.2.1 are found to be around 10%. While it turns out to be a significantly good value, it should not be forgotten that it has been obtained from the evaluation of a fairly small set of users. On the other hand, when evaluating the MAPE resulting from the models of the total users of the system an average of **15.42%** is obtained. Given that the total set of users is 92,794, this value is an indicator that the performance of the part of obtaining the statistical models is satisfactory. This is due to the fact that even with the large number of managed users, the average MAPE is set at a relatively low value.

In order to support the above statement, the frequency histogram of Figure 6.2-6 is presented, which allows to observe the ranges within which the MAPE of the total users are located. The range with the highest number of MAPE is between 10% and 11%, which contains 7,292 users. This, together with the analysis of the MAPE average, allows to validate the operation of this part of stage 2, due to its satisfactory performance in the modeling of the consumption profiles of the users of the system.

6.2.3 Intelligent correction of the system users models

For each user of the system, the correction is done by implementing a set of neural networks independent of each other. These networks are trained to be able to explain the ***differences*** between the user's actual consumption profile and its corresponding fitted statistical model. A metric based on the linear correlation coefficient to evaluate the performance of each network is used, the one with the best performance is selected to obtain the definitive ***differences*** that must be added to the statistical model in each period to correct the fitting of this for the user under study.

6.2.3.1 Experiment to select the best performing network

In Chapter 5 the reason for the selection of the neural networks as a technique to implement the second part of this stage was explained. Where the main idea is to

take advantage of the high variability that this technique presents in its output to explore a wide number of solutions. The above in order to find a neural network that is able to correctly generalize the behavior of the ***differences*** in each period of the study window, starting from the relations and interactions between the inputs of this second part of the stage (exogenous variables). It was necessary then to carry out an experiment that allows to test a varied number of networks until finding the one with the desired behavior.

The three factors that can be varied for the execution of this type of networks are the training algorithm, the number of neurons of the hidden layer and the number of networks tested. The training algorithm was defined as fixed, selecting the one that takes longer in execution but yields better results (Bayesian regularization). The number of neurons in the hidden layer remained fixed, this was obtained by means of a heuristic (see Section 5.3.5.2), which allows to avoid overfitting and underfitting. Therefore, the only factor that was varied was the number of trained networks.

To carry out the experiment, a set of neural networks are trained, all with the same configuration. Although all have the same configuration, the high variability of the technique allows obtaining a diverse set of outputs. It was defined that fifty (50) equal networks would be trained for each user.

In the previous chapter, the performance metric for the evaluation of the implemented neural networks (see Section 5.3.5.3) was defined, which is to obtain the linear correlation coefficient for each data set (training, validation and testing). These coefficients allow to analyze the performance of a network for each particular set. The evaluation of the total performance of said network is analyzed by computing the square sum of the three previous coefficients, which corresponds to the performance metric selected.

Next, the process of selecting the best network for a system user is exemplified. Because this part of the stage analyzes each user individually, this process is repeated exactly the same for all users of the system. Therefore, the example presented is sufficient and it is not necessary to show the process for the 92,794 users.

Figure 6.2-7 shows the linear correlation coefficients of the datasets for each of the fifty networks implemented during the analysis of the sample user. At first sight, it is not trivial to identify the network with the best overall performance. To achieve this, Figure 6.2-8, which contains the graph of the performance metric for each network is presented. The best network is the one where the highest point of the curve of said figure occurs, which has been indicated with a ***red*** dotted line. The maximum value mentioned occurs in network number 37, which is selected to perform the intelligent correction of the statistical model of the sample user.

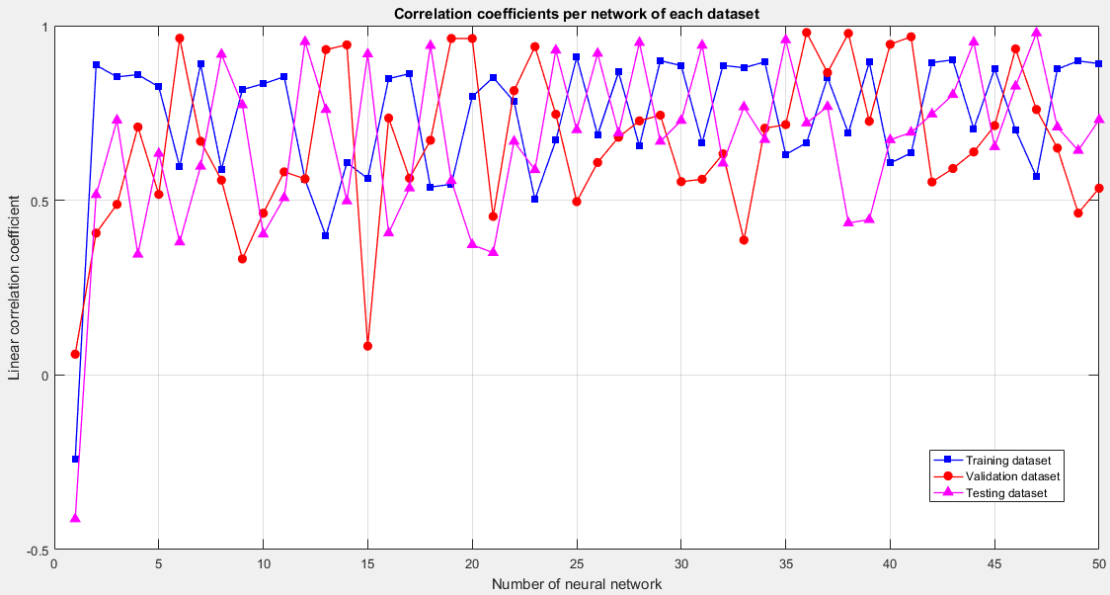


Fig 6.2-7. Correlation coefficients of datasets for each network.

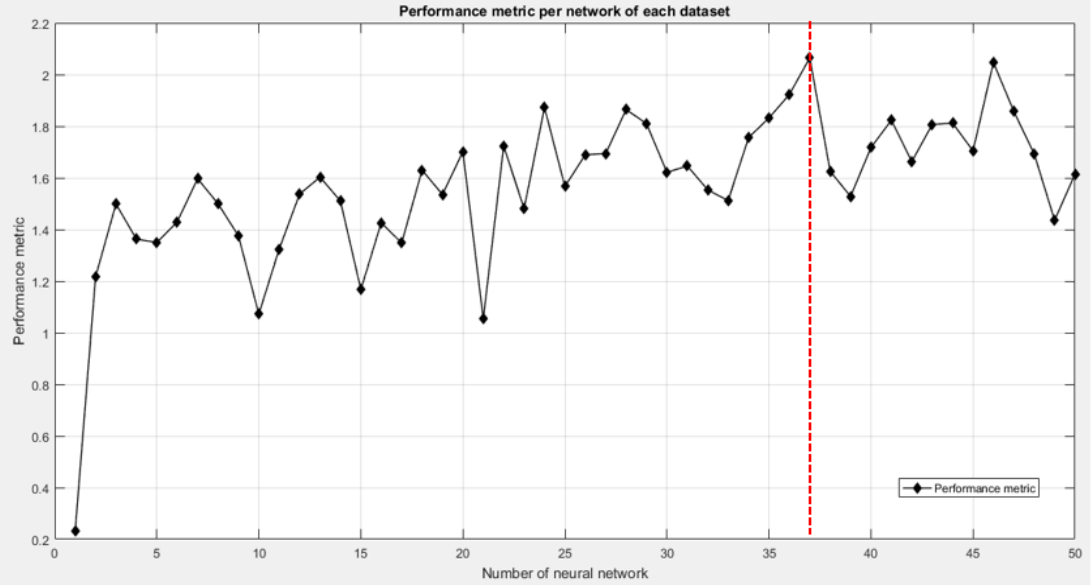


Fig 6.2-8. Performance metric for each network.

6.2.3.2 Obtaining the corrected models of each user

During the execution of this part of the stage for a user, after selecting the best network, it is executed to perform the intelligent correction of the statistical model obtained for that user. The execution of the network allows obtaining the ***differences*** that must be added in each period to the statistical model to improve the fitting made.

As this process occurs the same for the 92,794 users, in order to demonstrate the operation of this second part of stage 2 of the system, a sample of four example users has been selected. The results of the correction for these users are shown in the following figures. It should be noted that the selected examples correspond to users where the statistical model does not present a good performance, which makes it possible to notice the action of the intelligent correction.

Selected user 1

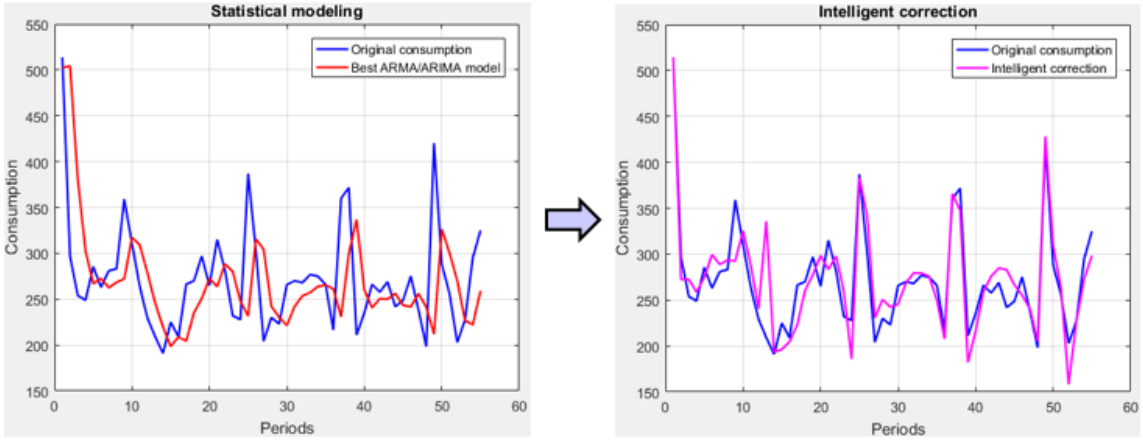


Fig 6.2-9. Intelligent correction of selected user 1.

Selected user 2

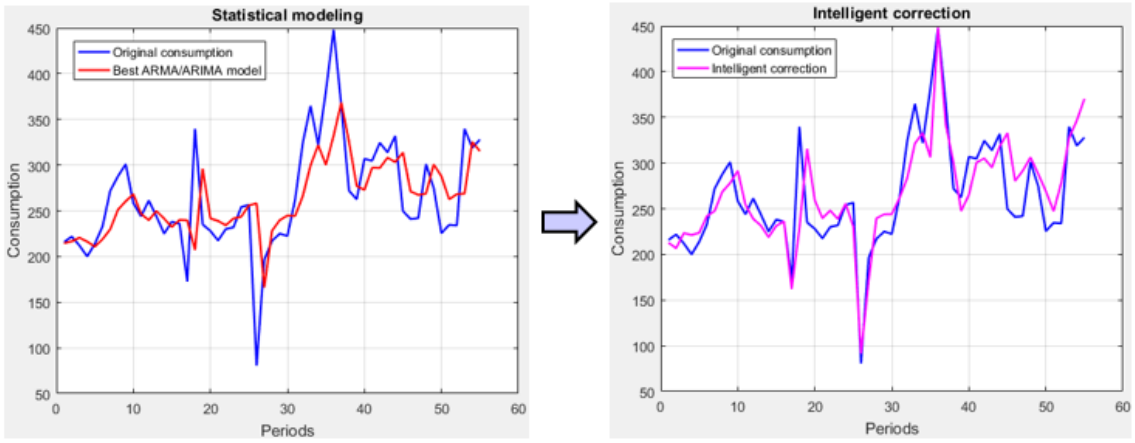


Fig 6.2-10. Intelligent correction of selected user 2.

Selected user 3

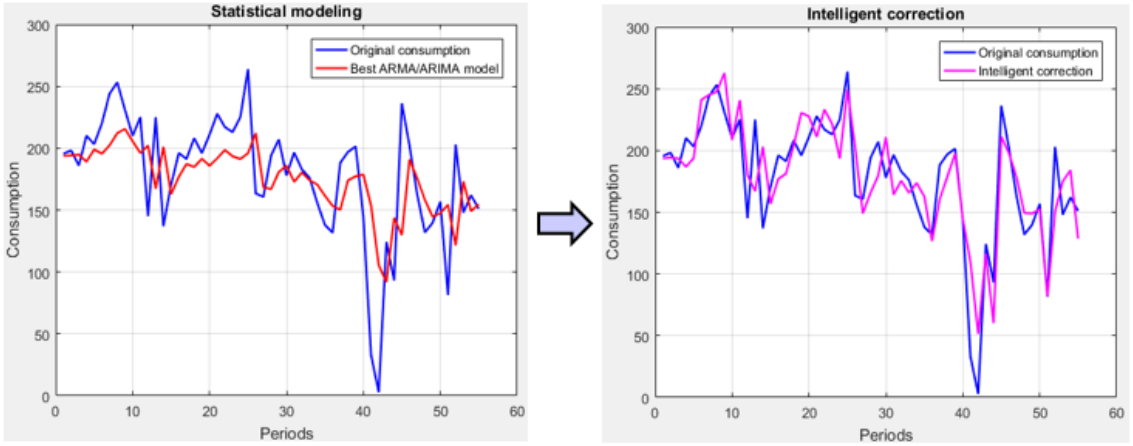


Fig 6.2-11. Intelligent correction of selected user 3.

Selected user 4

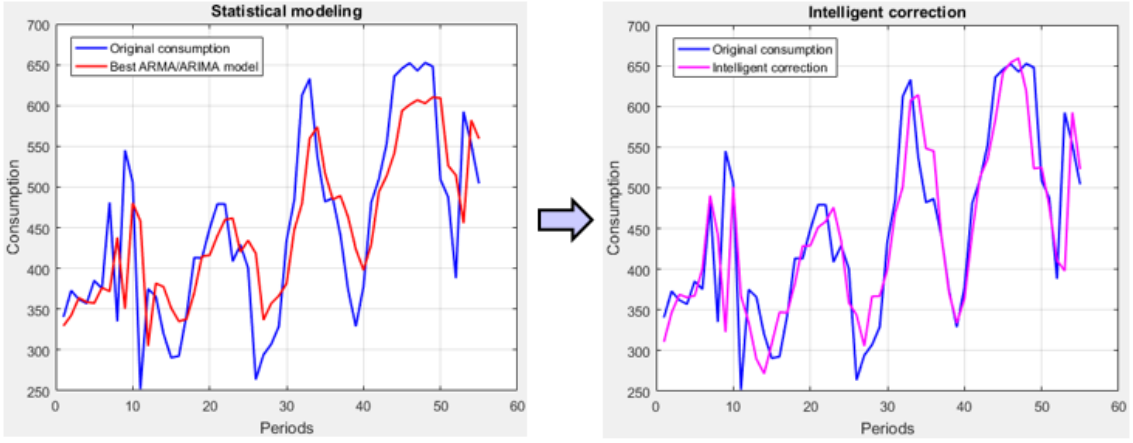


Fig 6.2-12. Intelligent correction of selected user 4.

Table 6.2-2 presents the MAPE obtained for each test sample before and after the intelligent correction. It is evident that for all cases the intelligent correction improves the adjustment made to the user's consumption profile.

Performance of correction		
Example user N°	**Stat.MAPE (%)**	**Correc.MAPE (%)**
1	16,22	8,03
2	14,31	9,31
3	89,70	46,01
4	13,40	9,55

Table 6.2-2. MAPE before and after intelligent correction.

In the previous chapter, when describing the operation of this part of stage 2, it was mentioned that the purpose was to make an ***attempt*** to correct the statistical model obtained. The above is due to the fact that, although the previous examples provide a notion of the correct functioning of the intelligent correction, this does not happen for all users. Figures 6.2-13 and 6.2-14 show examples of users in which the intelligent correction produces a decrease in the performance of their corresponding statistical model. The points marked in ***red*** on the corrected curve (***magenta*** line) indicate the values in which the correction has failed considerably.

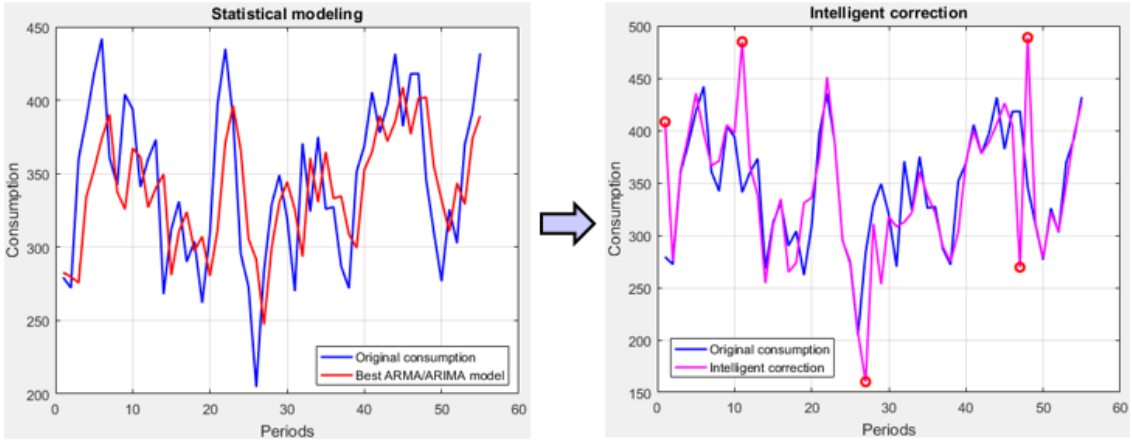


Fig 6.2-13. Decreased performance due to the intelligent correction, example 1.

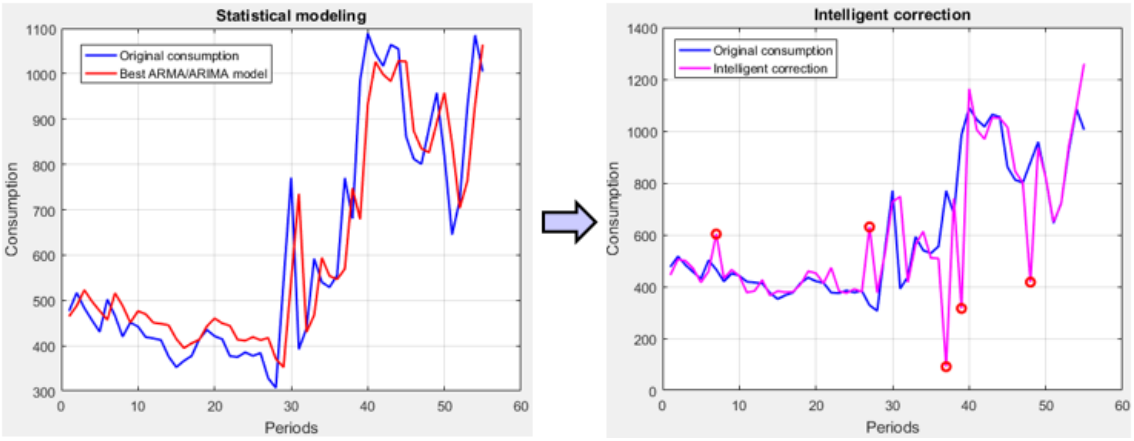


Fig 6.2-14. Decreased performance due to the intelligent correction, example 2.

Table 6.2-3 presents the MAPE obtained for the previous examples before and after the intelligent correction. The decrease in performance is verified as there is an increase in the MAPE of the statistical fittings.

Performance of correction		
Example user N°	**Stat.MAPE (%)**	**Correc.MAPE (%)**
1	11,71	12,33
2	12,84	14,19

Table 6.2-3. MAPE before and after the intelligent correction.

Similar to the case of the first part of this stage, the graphics of the MAPE distribution and the frequency histogram are presented, which are used for the evaluation and validation of the performance of the intelligent correction as a whole.

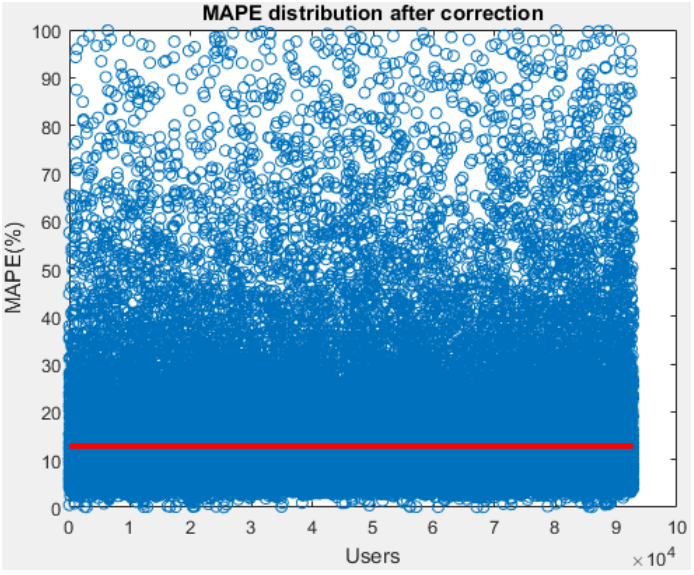


Fig 6.2-15. Distribution of MAPE after correction.

In Figure 6.2-15, the ***red*** line corresponds to the average of the MAPE obtained after the execution of the intelligent correction on all users of the system (92,794). The value of this average for this part is **12.83%,** which shows that in general, the execution of the correction manages to improve the performance of the fittings obtained in the first part. This is supported by the decrease obtained in the overall average of the system's MAPE, which dropped from 15.42% to 12.83%.

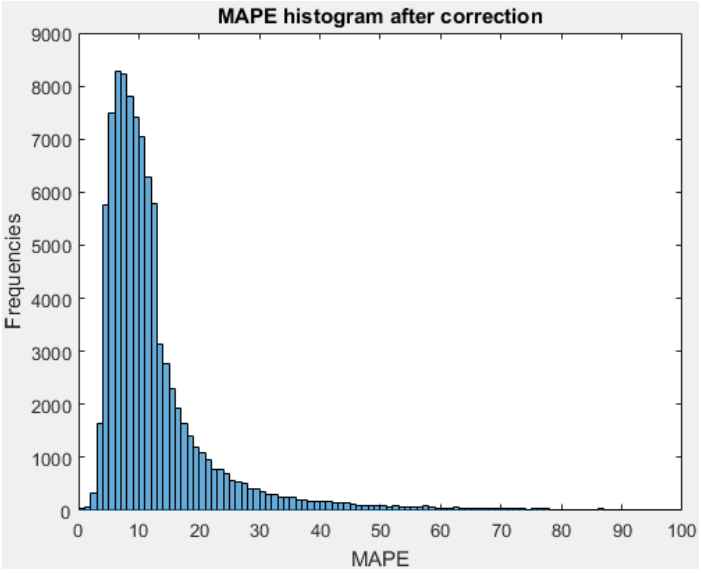


Fig 6.2-16. Histogram of the MAPE of the system after the intelligent correction.

The histogram in Figure 6.2-16 presents the same shape as that obtained for the performance evaluation of the first part (statistical modeling) (see Figure 6.2-6). However, it is possible to observe that the MAPE histogram of the correction has shifted to the left. This is consistent with the decrease in the overall MAPE average explained above. The range with the highest number of elements is between 6% and 7%, which contains 8277 users.

The analysis carried out allows to validate the performance of the second part of stage 2 of the system. This is due to the fact that this part seeks to make a correction of the statistical models obtained for each user in the first part, therefore, a decrease in the MAPE average and the histogram shift to the left

(lower MAPE values) allow to confirm that it fulfills the function of correction raised.

6.2.4 Selecting the prediction for each user

It was explained that the output of this stage is the ***consumption deviation*** of each user of the system, i.e., the difference between the closest prediction and the actual consumption of the predicted month for each of them. It should be kept in mind that after the execution of the stage two predictions have been made for each user, the one of the statistical model and the one corrected by the neural network. Although it was verified that the ***intelligent correction*** part works correctly, it was evident that in all users the performance is not improved, therefore, in all cases the corrected prediction will not be more accurate. The above represents the reason why it was defined that the comparison of both predictions with the actual value would be made to select the closest one.

After the execution of the complete stage, it was obtained that the number of users for whom the intelligent correction improves the statistical fitting (decreases the MAPE) is of 76,451. This corresponds to 82.38% of the total users of the system. However, when comparing the predictions, the situation changes. Figure 6.2-17 presents a ring diagram showing the number of users for whom the statistical prediction was finally used and the number of those for which the corrected prediction was used.

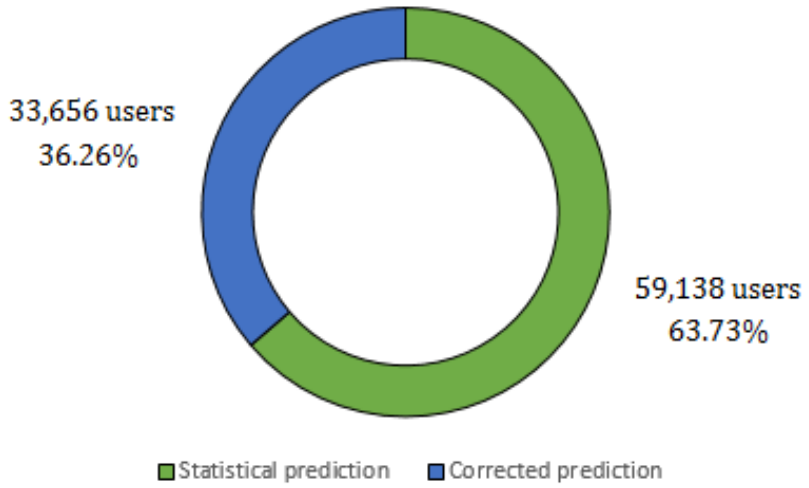


Fig 6.2-17. Ring diagram with the results of the comparison of predictions.

It is possible to observe that in spite of the good performance of the intelligent correction, only in a 36.26% it manages to make a more accurate prediction. Nevertheless, the result obtained is satisfactory. This is due to the fact that it is possible to improve 36.26% of the predictions using the intelligent correction, which would not have been possible using only the statistical predictor.

Due to the independent validation of the two parts of this stage, it has inherently been generally validated the stage 2 of the proposed intelligent system.

6.3. Tests and results of stage 3: Fraudulent user detection

Of the three stages that compose the system, although this is the most important, it also turns out to be the one whose implementation is simpler. It is the most important because it is where the classification is carried out to detect fraudulent users, which is the main objective of the proposed intelligent system. It is also the easiest to implement because its functional structure is essentially[3] given by the computational intelligence technique for classification.

From the above it can be said that once the sample dataset for the training is available, the functionality of the stage consists solely in the execution of the intelligent classification algorithm. Therefore, the tests were oriented towards the selection of the classification technique to be used and the configuration of the parameters of the same to obtain the best possible classification.

6.3.1 Selection of the intelligent classification technique

In the previous chapter it was mentioned that three techniques were considered for the implementation of this stage. These techniques are artificial neural networks, support vector machines and random forests; of which the one that presented the best performance in the solution of the problem of classification of users in fraudulent and non-fraudulent would be selected. An experiment was executed to compare the three techniques, where the evaluation of the performance of these was carried out by the most commonly used method to

[3] This expression is used to emphasize that the functionality of the stage is described only by the intelligent classification algorithm, however, at the structural level it also has the conditioning and obtaining output examples blocks (see Chapter 5, Section 5.4 .3.3).

evaluate classification algorithms, which is known as K-fold crossvalidation (Liu, Refaeilzadeh & Tang, 2008).

To describe it in a quick way, K-fold crossvalidation divides the total set of training data into ***K*** equal (or roughly equal) groups. ***K***-1 groups are used to train the algorithm and the remaining group is used for testing, the algorithm is executed and the results are recorded (performance). Next, ***K***-1 groups are used again to train the algorithm a second time and the remaining group is reserved for testing. It should be kept in mind that in this second run, the group that was first used for testing must be part of the ***K***-1 used for the training, with a different one being used for testing. The process is repeated until the ***K*** groups have been used for testing, i.e., the technique is executed a total of ***K*** times. The overall performance of the classification algorithm is the average of the performances that were obtained for each of the ***K*** testing groups.

Figure 6.3-1 shows an example of K-fold crossvalidation for a ***K*** of 10. It is observed that in the first round the dataset is divided into ten groups, nine are used for training and one for validation. In the second round, the group that was previously validation became part of the training set, taking a different one for validation. The process is repeated in the remaining rounds until the ten groups have been used for validation. The final performance is the average of the performances obtained on the validation group in each round.

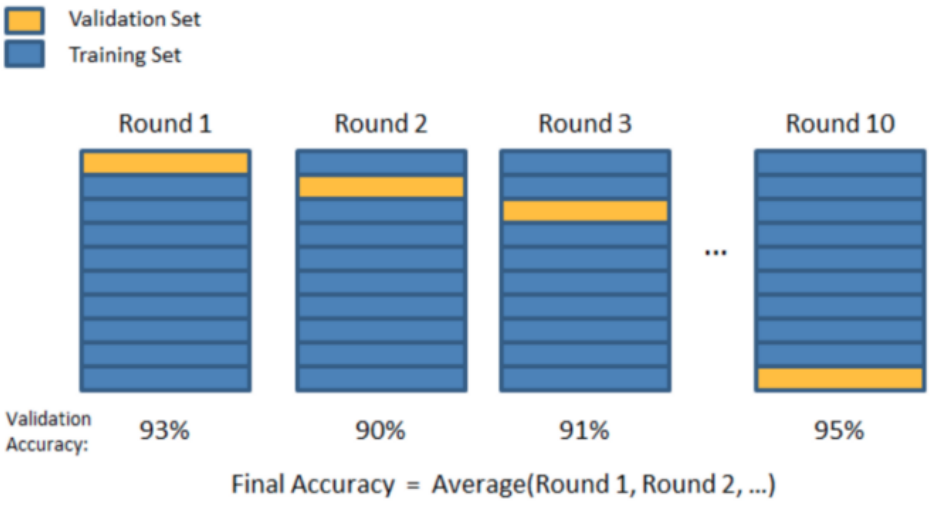


Fig 6.3-1. Example of K-fold crossvalidation with a K of 10. (www.medium.com)

6.3.1.1 Experiment to select the classification technique

In Section 5.4.4 it was mentioned that the number of periods (months) for training the detector is six (November 2014 to April 2015). The set of examples that results from examining these periods consists of 4640 users. The previous amount is considerably smaller than the 92.794 of the total users of the system. However, this is explained because the sample base is constructed from the variables of ***macrometers*** (non-fraudulent users) and ***inspection results*** (fraudulent users) (see Chapter 5, Section 5.4.3.1).

The commercialization company of the region carries out between 20,000 and 30,000 monthly customer inspections, this number is total for the entire Caribbean region. It is possible to estimate that the number of monthly inspections carried out on users belonging to the ***Atlantico Norte*** delegation is well below the total number presented. In addition, it should be remembered that the total users of the system are the product of a series of filters, therefore, of the inspections carried out in ***Atlantico Norte***, those that fall on valid users of the system will be even smaller. All of the above allows explaining the reason for obtaining only 4640 users after the review of the six periods intended for the training of the detector.

One of the conditions for the correct training of a classification algorithm is to guarantee the homogeneity of the training base, that is, to have equal number of examples of each class. This in order that the algorithm can perform a good generalization, since in highly unbalanced training sets (appreciable difference between the number of examples for each class) tends to present a bias in learning towards classes with more examples.

From the above it is possible to infer that of the 4640 training examples, 2320 are fraudulent users and the remaining 2320 are non-fraudulent. On average, 750 users are given monthly to feed the bank of examples of training, approximately 375 of each class (fraudulent and non-fraudulent).

The experiment that was performed for the comparison of the three techniques was a K-fold crossvalidation with a ***K*** of 10, which was executed three times (1 time for each technique). The complete base of 4640 users was used for the

experiment, which means that 10 groups each of 464 users were obtained for crossvalidation, using in each round 9 for training (4176 users) and 1 for validation (464 Users).

The network topology, number of neurons of the hidden layer and the training algorithm were fixed for the configuration of the neural network. The topology was a two-layer feed-forward, the number of neurons was obtained by the expression presented in Section 5.3.5.2 and the training algorithm used corresponds to the best performance (in classification problems) with a longer training time. The parameter to be varied was the number of trainings of the network, that is to say, the same network is trained several times to find the one of better results. To configure the SVM the Kernel function was fixed, using the most common for classification problems. Gamma coefficient and cost were varied. To configure the random forest, the only parameter is the number of trees, which was subject to variation.

A set of pilot tests were performed to determine the values of the variable parameters of each technique that allow to obtain ***good***[4] results on the training dataset, which were used for the experiment of comparison of the techniques. Table 6.3-1 compiles the values of the fixed and variable parameters of each technique used for the comparison.

Parameters of techniques	
Two-Layer feed-forward ANN	
N° of hidden layer neurons	35
Training algorithm	Resilient backpropagation
N° of trainings of the ANN	50
Support vector machines	
Kernel Function	Radial Basis
Gamma	0,01
Cost	10
Random Forests	
Number of trees	50

[4] The word ***good*** stands out to indicate that the pilot tests were aimed at finding a neighborhood of the parameters of each technique in which good performances were obtained, however, they were not so rigorous as to find the best values of these parameters.

Table 6.3-1. Parameters of techniques for crossvalidation.

Table 6.3-2 compiles K-fold crossvalidation with ***K*** of 10 results for the three techniques, which show the training and testing performances for each of the 10 rounds. The performance represents the success rate (percentage of hits) of each technique, i.e., in what percentage of the times indicated as fraudulent a fraudulent user and indicated as non-fraudulent a non-fraudulent user.

K-fold Crossvalidation, *K* = 10						
Round	Training performance (%)			Testing performance (%)		
	ANN	SVM	Random forests	ANN	SVM	Random forests
1	78,90	86,09	94,84	74,50	76,18	83,11
2	79,20	85,99	93,15	77,90	73,70	82,34
3	78,50	85,93	94,94	79,50	74,13	80,49
4	78,40	86,93	95,08	78,40	72,95	82,11
5	78,80	86,63	94,84	78,20	72,19	81,57
6	78,30	86,01	94,76	80,90	74,56	80,38
7	78,90	85,82	94,92	76,70	74,67	82,13
8	79,50	86,90	95,12	78,80	71,33	82,00
9	78,70	86,66	94,70	80,30	70,68	81,03
10	79,60	86,42	94,76	77,90	74,89	80,19
Average	**78,88**	**86,34**	**94,71**	**78,31**	**73,53**	**81,54**

Table 6.3-2. Results of K-fold crossvalidation to compare techniques.

The values highlighted in ***yellow*** indicate the technique with the highest average performance for each case (training and testing), which shows that for both cases, random forest is the technique with the best average performance. Therefore, it was the technique selected for the implementation of the classifier of stage 3 of the system (detection). Additionally, for both cases the graphs showing the performances of the techniques for each round of K-fold crossvalidation are presented.

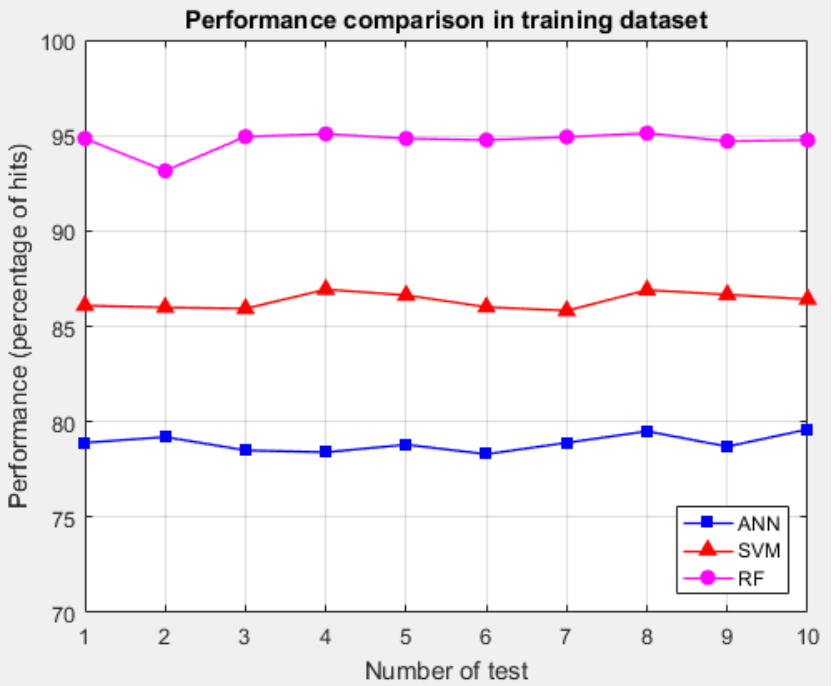


Fig 6.3-2. Results per round of K-fold crossvalidation for training.

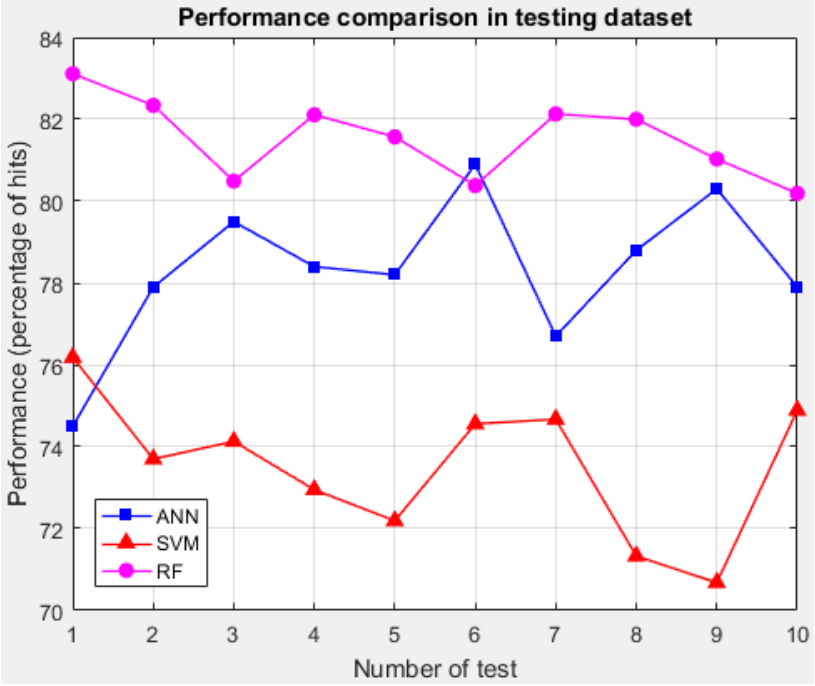


Fig 6.3-3. Results per round of K-fold crossvalidation for testing.

Figures 6.3-2 and 6.3-3 provide a visual check of what is presented in Table 6.3-2, where it is possible to see the superior performance of random forests. The above

because the performance curves for this technique are located above the curves of the other two techniques compared, for both training and testing.

6.3.2 Selection of the number of trees

Once selected random forests as the technique to be used, it was necessary to find the values of the parameters of this one that allow to obtain the best performance of the classification. It was mentioned previously that for this technique the only parameter that can be varied is the number of trees of the model. Therefore, an experiment based on K-fold crossvalidation was performed again to determine the number of trees that allow to obtain the best performance.

K-fold crossvalidation was used with $\boldsymbol{K}$ of 5, the decrease in the value of $\boldsymbol{K}$ is justified in the low variability of the output of the technique (see Figures 6.3-2 and 6.3-3). The total training base (4640 users) was used to carry out the experiment. This means that 5 groups each of 928 users were obtained for crossvalidation, using in each round 4 training (3712 users) and 1 for validation (928 users). The number of trees to test was from 10 to 100 in steps of 10.

A total of ten K-fold crossvalidations were performed with $\boldsymbol{K}$ of 5, this being due to the fact that for each value of the number of trees it was necessary to obtain the general performance of the technique. Table 6.3-3 compiles the average performance results for each number of trees in training and testing cases. Additionally, these results are shown graphically in Figure 6.3-4.

K-fold Crossvalidation, K = 5		
N° of Trees	**Average Training performance (%)**	**Average Testing performance (%)**
1	93,52	82,87
2	94,68	81,47
3	95,41	81,36
4	95,54	82,00
5	95,54	82,00
6	95,60	82,87
7	95,35	81,47
8	95,52	82,44
9	95,70	81,36
10	95,35	81,57

Table 6.3-3. Results of K-fold crossvalidation for each number of trees.

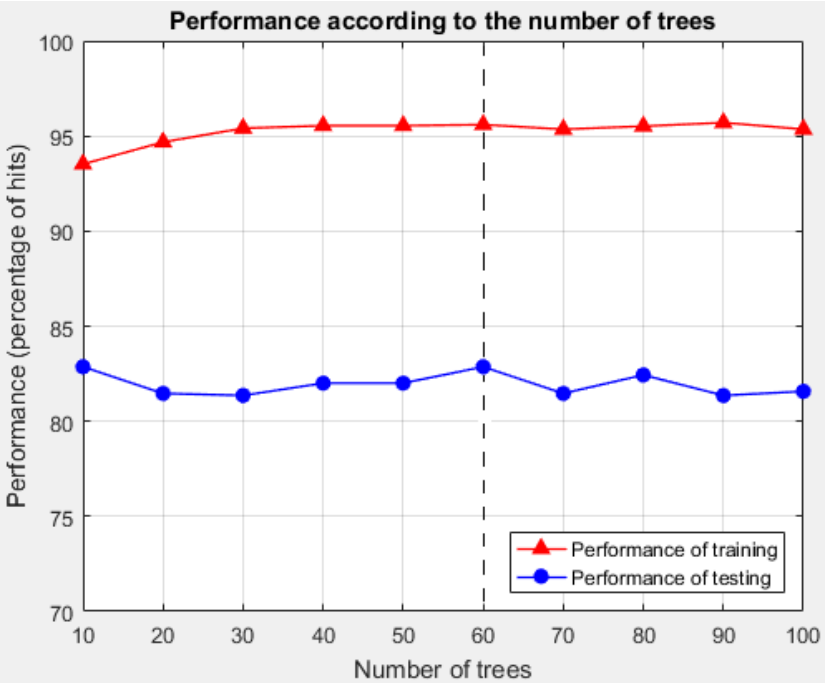


Fig 6.3-4. Results of K-fold crossvalidation for each number of trees.

From the presented results the value of 60 is selected for the parameter of the number of trees of the model. The best average performance of the training is the model of 90 trees with 95.7%, the model of 60 trees has the second best average performance in the training with 95.6%. However, in the testing scenario the 90-tree model has the lowest average performance with 81.36%, while the 60-tree model has the highest average performance with 82.87%. The small difference in the training case and the big difference in the testing makes the 60-tree model the one with the best performance.

6.3.3 Fraudulent user detection

From the experiment of selection of the number of trees is obtained the model of 60 trees trained with the training dataset of 4640 users (6 periods). The stage 3 and system validation was performed through the execution of the detector in the 3 periods reserved for testing (May 2015 to July 2015). It should be remembered that the execution of the system is monthly, therefore, each period will be tested independently.

The extraction of the users from each of the test periods allowed to obtain three sets of data, which are summarized in table 6.3-4. It is possible to observe that although some of these are not perfectly homogeneous, the difference is minimal and does not generate any affectation in the execution of the system.

Reserved testing datasets			
Test period	**Fraudulent users**	**Non-Fraudulent users**	**Total users**
1 - May 2015	494	513	1007
2 - June 2015	453	461	914
3 - July 2015	549	549	1098

Table 6.3-4. Datasets reserved for system validation.

Section 5.4.5.2 described the performance metric used in machine learning for classification problems, which is the confusion matrix. However, this is nothing more than a graphic and detailed representation of the success rate (percentage of hits) of the classifier (as used in previous sections). The confusion matrix was used to evaluate the performance of the direct application of the system in each of the testing sets.

It should be noted that the three sets (periods reserved for testing) were not used during training, so they were completely new and unknown to the system. The results of these sets were known and used to compare them with those obtained from the execution of the system, the above in order to evaluate the performance of the detection. The execution of the detector yielded the following results.

First testing period (May 2015)

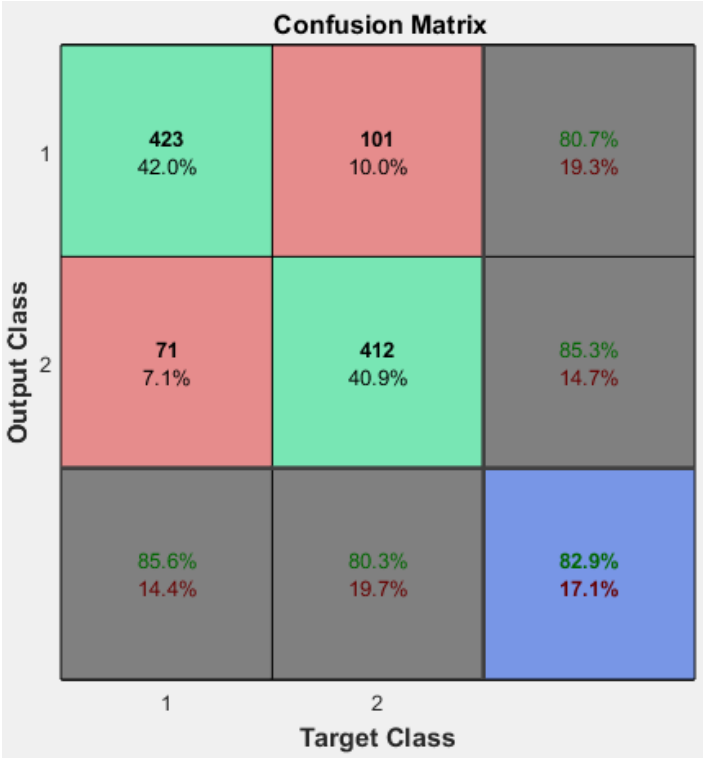


Fig 6.3-5. Confusion matrix for the first testing period.

Class 1 is for fraudulent users and Class 2 for non-fraudulent users. It is possible to observe that the performance (percentage of hits) of the classification for the first period is of **82.9%.** Of the 494 fraudulent users the detector correctly identified 423, failing 71 users. Of the 513 non-fraudulent users the detector correctly identified 412, failing 101 users.

Second testing period (June 2015)

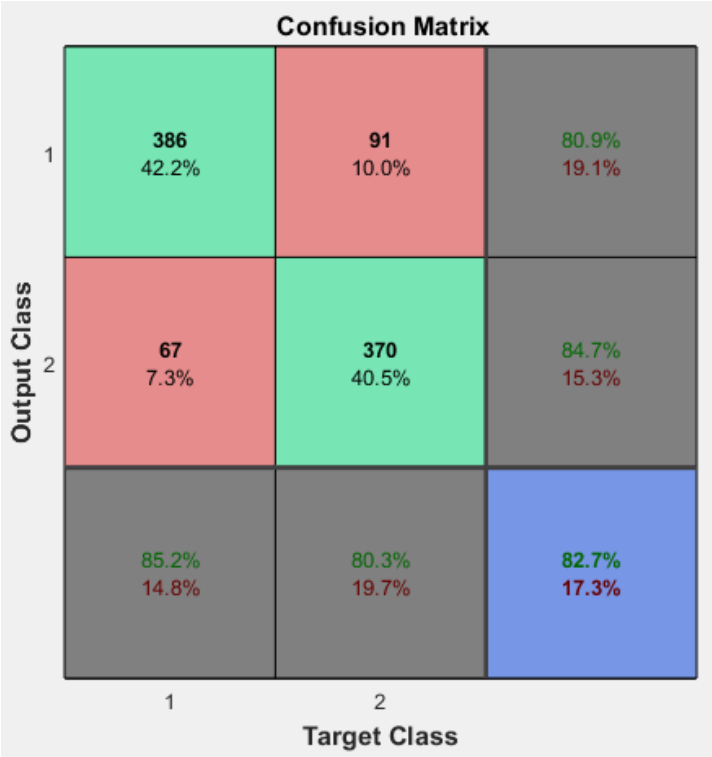


Fig 6.3-6. Confusion matrix for the second testing period.

Class 1 is for fraudulent users and Class 2 for non-fraudulent users. The performance (percentage of hits) of the classification for the second period is **82.7%**. Of the 453 fraudulent users the detector correctly identified 386, failing 67 users. Of the 461 non-fraudulent users the detector correctly identified 370, failing 91 users.

Third testing period (July 2015)

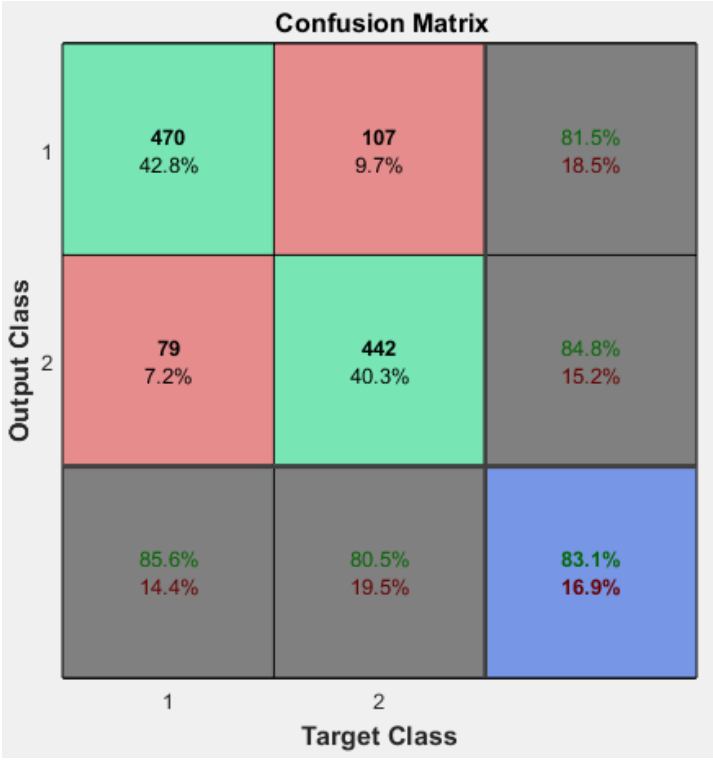


Fig 6.3-7. Confusion matrix for the third testing period.

Class 1 is for fraudulent users and Class 2 for non-fraudulent users. The performance (percentage of hits) of the classification for the third period is **83.1%**. Of the 549 fraudulent users the detector correctly identified 470, failing 79 users. Of the 549 non-fraudulent users the detector correctly identified 442, failing 107 users.

6.3.4 Overall performance and system validation

From the results obtained for the three periods reserved for testing, the system presents an average performance of **82.9%**, successfully classifying fraudulent and non-fraudulent users from the variables used to characterize them (including the two reinforcement inputs). This places the system in a good position within the group of works reviewed in Chapter 3, because it surpasses the performance of most of these.

Beyond the comparison with other works, the main point of comparison is in the place of action of the system. It is known that the commercialization company of the Colombian Caribbean region has a computer tool for the management of non-technical losses. Although the structure and functionality of this tool are private and therefore unknown, the company publishes in its annual report for 2015 that the result of the execution of the tool for energy recovery campaigns is 68%. In spite of only having three periods (three months of 2015) to validate the proposed system, the results obtained show that it has low variability in its output, which allows to expect a similar behavior in relation to the execution with new periods. Therefore, it can be affirmed that the proposed system has a superior performance than the tool used by the commercialization company of the region.

From the satisfactory results obtained and the analyzes performed, the operation of the proposed intelligent system as a tool for the detection of fraudulent users is validated.

Chapter 7

Conclusions and Main contributions

This chapter summarizes the main conclusions arisen of the analysis and discussion of the results reported in this work. It also reviews the dissertation's scientific contributions.

7.1. Conclusions

The results obtained from the execution of the proposed intelligent system allowed to identify the outstanding characteristics of this and of each one of the stages that compose it. It was possible to evaluate the performance of the system in achieving the main objective for which it was designed, which is the detection of users with fraudulent behavior. With respect to the above, it was possible to establish that the system has an average overall performance of **82.9%** in identifying the class to which each evaluated user belongs. However, because this is composed of multiple stages, it is necessary to conclude about each of the parts that have made possible to achieve the performance mentioned.

- Many variables were selected to characterize the electrical energy consumption behavior of the users under study. However, the objective is not to select a large number of variables, but to identify those that provide the greatest amount of information to complete the characterization. In this played a key role to have counted with the experience of people who work in the companies of commercialization of energy, because it led to the selection of the variables needed.
- Although the clustering of consumption profiles was used in several works of other authors reviewed in Chapter 3, the one implemented in this proposal was optimized by the use of a genetic algorithm. Allowing to find

the configuration with the best performance of the grouping, which improves the results obtained from this grouping process.

- Essentially, the approach of stage 2 was oriented to obtain a correct prediction of the consumption of the clients. This was carried out by means of a statistical block that adjusted the best model of each user, from which an average general MAPE of 15.42% was obtained. Then an intelligent correction of the statistical models was performed, this allowed to reduce the average general MAPE to 12.83%. Leading to a performance increase in the prediction of 36.26%.
- Classification problems require the use of a technique capable of recognizing the patterns present in the data. To ensure the correct solution to the problem of detecting fraudulent users, several techniques were compared and the one that presented the best performance, which was random forests, was selected. This led to the achievement of the mentioned performance of 82.9%, which was considered satisfactory and allowed to validate the operation of the system.

7.2 Main Contributions

This thesis presents a computational tool oriented to the management of non-technical losses of electrical energy, specifically aimed at the detection of users with fraudulent connections. In this case an intelligent system composed of three stages is proposed, so that a given set of users can be classified as fraudulent and non-fraudulent based on the variables used to characterize them and the overall execution of the stages of the system. The approach proposed in this research presents satisfactory performance within those reviewed in the state of the art and allows to obtain better results than the tool used by the commercialization company in the geographical space of study of the system.

The contributions made by this thesis to the problems associated with non-technical losses that are the product of fraudulent connections are:

- *An intelligent approach for the detection of fraudulent users using a system composed of a stage of grouping similar consumption profiles, an intelligent-*

corrected consumption profile modeling and prediction stage, and a classifier to carry out the desired detection.

- *The algorithms that constitute the functional structure of each of the stages of the system. These compose a tool with a satisfactory performance in the detection of fraudulent users.*

References

1. **(Navani, Sharma & Sapra, 2012)** Navani, J. P., Sharma, N. K., Sapra, S. (2012). Technical and non-technical losses in power system and its economic consequence in Indian economy. International Journal of Electronics and Computer Science Engineering, 1(2), 757-761.

2. **(Suriyamongkol, 2002)** Suriyamongkol, D. (2002). Non-technical losses in electrical power systems. United States. Ohio University.

3. **(Unidad de Planeación Minero Energética, 2011)** Unidad de Planeación Minero Energética. (2011). Plan Preliminar de Expansión de Referencia Generación – Transmisión 2011-2025, Capítulo 2, p. 25-27. Ministerio de Minas y Energía, Colombia.

4. **(Czernichow, Muñoz & Sanz-Bobi, 1998)** Czernichow, T., Muñoz, A., Sanz-Bobi, M. A. (1998). System for detection of abnormalities and fraud in customer consumption. 12th Conference on the Electric Power Supply Industry. Pattaya, Thailand.

5. **(Cabral, Martins, Pinto. A, & Pinto. J, 2008)** Cabral, J. E., Martins, E. M., Pinto, A. M., Pinto, J. O. (2008). Fraud detection in high voltage electricity consumers using data mining. Brazil. University of Mato Grosso do Sul.

6. **(Biscarri. F, Biscarri. J, Guerrero, León, Millán & Monedero, 2010)** Biscarri, F., Biscarri, J., Guerrero, J., León, C., Millán, R., Monedero, I. (2010). Increasing the efficiency in non-technical losses detection in utility companies. Spain. University of Seville.

7. **(Hashim, Hussien, Mohamad, Pok & Yak, 2007)** Hashim, A. H., Hussien, Z. F., Mohamad, A. M., Pok, H. L., Yap, K. S. (2007). Abnormalities and fraud electric meter detection using hybrid support vector machine and modified genetic algorithm. 19th International Conference on Electricity Distribution.

8. **(Nagi, Yap, Tiong, Ahmed & Mohammad, 2008)** Nagi, J., Yap, K. S., Tiong, S. K., Ahmed, S. K., Mohammad, A. M. (2008). Detection of abnormalities and electricity theft using genetic support vector machines.

9. **(Nizar & Dong, 2009)** Nizar, A. H., Dong, Z. Y. (2008). Identification and detection of electricity customer behavior irregularities.

10. **(Figueiredo, Muniz, Tanscheit & Vellasco, 2009)** Figueiredo, K., Muniz, C., Tanscheit, R., Vellasco, M. (2009). A neuro-fuzzy system for fraud detection in electricity distribution. Brazil. Pontifical Catholic University of Rio de Janeiro.

11. **(Carmona, Nunes, Saavedra & Silva, 2011)** Carmona, O., Nunes, A., Saavedra, O., Silva, E. (2011). Detection and identification of abnormalities in customer consumptions in power distribution systems. IEEE Transactions on Power Delivery, 26(4), 2436-2442.

12. **(Kundur, 1994)** Kundur, P. (1994). Power system stability and control, Chapter 1, p. 5-8. McGraw-Hill.

13. **(Ley N° 142, 1994)** Ley N° 142. Diario Oficial número 41,433. Colombia, 11 de Julio de 1994.

14. **(Ley N° 286, 1996)** Ley N° 286. Diario Oficial número 42,824. Colombia, 5 de Julio de 1996.

15. **(Elexon, 2013) Elexon. (2013).** Load profiles and their use in electricity settlement, p. 1-2. Retrieved from https://www.elexon.co.uk /wp-content/uploads/ 2013/11/load_profiles _v2.0_cgi.pdf.

16. **(Derivex, 2010)** Derivex. (2010). Caracterización del mercado eléctrico colombiano, p. 11.

17. **(Rios, 2013)** Rios, S. (2013). Guía para la detección de frauds en suministros de energía eléctrica en medición directa. Colombia. Universidad Tecnológica de Pereira.

18. **(Nilsson, 1998)** Nilsson, N. (1998). Introduction to machine learning, Chapter 1, p. 1-3; Chapter 9, p. 120-125. United States. Stanford University.

19. **(Anonymous, 2017)** Anonymous. (2017). Machine learning notes for beginners, Chapter 1, p. 2-3.